Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO DE FIN DE MÁSTER

Máster en Matemáticas

CÁLCULO VARIACIONAL EN EL ESPACIO DE WASSERSTEIN

Autor: Alberto Martín Heras

Tutor: Eustasio del Barrio Tellado

2024

Tabla de contenidos

Resumen

Numerosas aplicaciones recientes en análisis de datos se basan en la consideración del conjunto de observaciones como una realización con ruido de una medida de probabilidad. En este contexto es habitual el uso de métodos en los que se compara esta probabilidad observada con los elementos de una familia de distribuciones, buscando la que mejor se ajusta en el sentido de minimizar una distancia o divergencia. La métrica de Wasserstein, asociada el problema de transporte óptimo, es una de las opciones más estudiadas en tiempos recientes, por sus buenas propiedades de adaptación a la geometría de los datos [1]. Esto justifica el interés de estudiar los problemas de minimización de distancias de Wasserstein respecto a familias de probabilidades. En la práctica esta minimización requiere el desarrollo de algoritmos de tipo descenso de gradiente en el espacio de Wasserstein, lo que resulta factible gracias a su estructura pseudo-Riemanniana (ver [2]). En algunos casos el paso de gradiente se puede implementar de forma explícita (familias univariantes, familias elípticas, ver [3]). En otros casos es necesario recurrir a algún tipo de discretización. En este Trabajo de Fin de Máster se estudiarán las dos formas principales de discretización: Euleriana (discretización del espacio de referencia en celdas fijas y cuantización correspondiente de las probabilidades) o Lagrangiana (aproximación de las probabilidades mediante versiones empíricas). Se analizará las ventajas e inconvenientes de las dos aproximaciones en problemas de alta dimensión y se prestará atención a las posibles ganancias computacionales asociadas a la paralelización masiva de cálculos.

Palabras clave

Transporte óptimo, regularización entrópica, algoritmo de Sinkhorn, espacio de Wasserstein.

Abstract

Many recent applications in data analysis are based on considering the set of observations as a noisy realisation of a probability measure. In this context it is common to use methods in which this observed probability is compared with the elements of a family of distributions, looking for the best fit in the sense of minimising a distance or divergence. The Wasserstein metric, associated with the optimal transport problem, is one of the most studied options in recent times, due to its good properties of adaptation to the geometry of the data [1]. This justifies the interest in studying Wasserstein distance minimisation problems with respect to probability families. In practice this minimisation requires the development of gradient descent type algorithms in Wasserstein space, which is feasible thanks to its pseudo-Riemannian structure (see [2]). In some cases the gradient step can be implemented explicitly (univariate families, elliptic families, see [3]). In other cases it is necessary to resort to some kind of discretisation. In this Master Thesis we will study the two main forms of discretisation: Eulerian (discretisation of the reference space into fixed cells and corresponding quantization of the probabilities) or Lagrangian (approximation of the probabilities by empirical versions). The advantages and disadvantages of the two approaches in high-dimensional problems will be analysed and attention will be paid to the possible computational gains associated with the massive parallelisation of computations.

Keywords

Optimal transport, entropic regularization, Sinkhorn's algorithm, Wasserstein space.

Introducción

El problema de transporte óptimo fue propuesto por primera vez por Monge a finales del siglo XVIII. La formulación que propuso surgía de forma natural al considerar la forma óptima de transportar unos recursos desde una serie de puntos de partida hasta una serie de destinos para satisfacer unas demandas. El marco general en el que se establece el formalismo del problema de Monge es la teoría de la probabilidad, en particular se emplea el lenguaje de transporte de probabilidades para dotar de rigor a la formulación de Monge el cual es un problema de minimización. La principal dificultad que surge de este enfoque es que la función objetivo no es convexa, de hecho el conjunto de aplicaciones de trasporte no es convexo. Esta situación impide emplear la maquinaria de la optimización convexa para estudiarlo y lo convierte en un problema difícil de estudiar.

En la década de 1940, Kantorovich propuso una formulación más laxa del problema de transporte óptimo que tiene la ventaja de ser un problema de optimización convexa. Esta se basa en la búsqueda de planes de transporte óptimo en vez de aplicaciones de transporte óptimo. Para la formalización de los planes de transporte óptimo se emplean distribuciones conjuntas en un espacio producto. Esta nueva perspectiva del problema inicial permitió el desarrollo de una teoría muy potente llegándose a dar solución al problema de Monge en el caso del coste cuadrático en $\mathbb{R}^n$. El transporte óptimo ha tenido repercusiones importantes en áreas de la economía, logística y recientemente en aprendizaje automático. La aplicación del problema a este sector en particular se debe a la distancia de Wasserstein, la cual se define como un coste de transporte óptimo. Su utilidad radica en que permite comparar dos probabilidades. Esta situación ocurre con mucha frecuencia porque los datos de un modelo de aprendizaje pueden ser descritos por una probabilidad con soporte finito. La medida de discrepancia entre dos probabilidades que proporciona la distancia de Wasserstein es heredada del espacio subyacente en el que están definidas estas. Como se detallará en el trabajo, esta distancia tiene en cuenta la localización de las probabilidades y proporciona ventajas frente a otras medidas de discrepancia.

El problema de transporte óptimo pese a ser un problema de optimización lineal es costoso de resolver en el caso general y en su formulación discreta es un problema de programación lineal para el cual no se tienen algoritmos con capacidad de computación adecuada para aplicaciones de tamaño moderado. En este contexto, se plantea la introducción de una regularización entrópica a la formulación de Kantorovich con el fin de obtener un problema de optimización estrictamente convexo. Esta penalización entrópica a pesar de alterar el valor del coste de transporte entrópico, permite establecer resultados de unicidad. En el caso discreto se pueden apreciar aún más las ventajas de esta modificación, ya que permiten la obtención de la solución óptima mediante un método iterativo de punto fijo fácil de implementar. Este es el Algoritmo de Sinkhorn, el cual es responsable de la creciente popularidad del transporte entrópico en aplicaciones de aprendizaje automático. Esto se debe a que es una algoritmo paralelizable que solamente efectúa productos matriciales y operaciones elementales sobre vectores. Su uso ha permitido el empleo del coste de transporte óptimo y la distancia de Wasserstein como función de pérdida en numerosas aplicaciones como el tratamiento de imágenes.

En este trabajo se va a recorrer las distintas formulaciones del problema de transporte óptimo empezando por el enfoque de Monge. Seguidamente se expondrá la formulación de Kantorovich la cual resulta ser la adecuada para el estudio de la teoría de transporte óptimo. Se analizarán las principales propiedades del problema de transporte óptimo y se dará la definición de la distancia de Wasserstein. Esta es la que da nombre a los espacios de Wasserstein, los cuales son espacios métricos que transportan la distancia del espacio subyacente a distancias entre probabilidades.

El segundo capítulo del trabajo se centra en el estudio en el transporte entrópico. En él se demostrarán los resultados principales de la teoría y se describirá la relación de este con el transporte óptimo. Se probará la existencia y unicidad del plan de transporte entrópico y se introducirá la distancia de Wasserstein regularizada. Esta no sera más que la distancia de Wasserstein usual a la que se le sumará un término que jugará un papel de penalización entrópica. La importancia de esta distancia regularizada se verá en los capítulos posteriores.

En el Capítulo 3 se detallarán las propiedades del transporte entrópico en el caso discreto y se profundizará en el Algoritmo de Sinkhorn probando su convergencia. La descomposición de la solución óptima del problema

de transporte entrópico discreto tendrá especial relevancia en el cálculo efectivo de esta como en los resultados de diferenciabilidad que se darán al final del capítulo. En ellos, se estudiará la regularidad del coste de transporte entrópico discreto respecto de cada uno de sus argumentos, lo que será esencial para la implementación de métodos de minimización basados en el Algoritmo de Descenso de gradiente.

En el último capítulo se empleará el marco de los espacios de Wasserstein para el estudio de varios problemas variacionales, los cuales se formulan en relación a un coste de transporte óptimo. La estrategia que se seguirá con el fin de simplificar los problemas será considerar una versión regularizada de ellos y una configuración discreta. De esta manera, se obtendrán problemas que podrán ser estudiados con las técnicas que se han introducido en el Capítulo 3. Un paso necesario para el empleo de las simplificaciones propuestas será la discretización de probabilidades. En este trabajo se expondrán dos esquemas que permiten trabajar con una aproximación discreta de una probabilidad continua y se probará su comportamiento asintótico. En particular, se introducirá la discretización Euleriana y la Lagrangiana, las cuales aportaran un enfoque distinto a cada uno los distintos problemas variacionales que consideraremos.

Capítulo 1

Transporte Óptimo

En esta sección se va a introducir el problema de transporte óptimo en su formulación propuesta por Monge y se estudiará la relajación del problema introducida por Kantorovich. La segunda formulación resuelve las dificultades que aparecen en el planteamiento de Monge.

En esta sección se empleará $\mathcal{X}, \mathcal{Y}$ para denotar dos espacios medibles arbitrarios y se seguirá el Capítulo 4 de [4].

1.1. Motivación

Se tiene un recurso que se quiere transportar en su totalidad desde una serie de orígenes hasta unos destinos de forma que se supla la demanda de este recurso en cada uno de estos destinos. Se asume que la cantidad total de recurso que se debe transportar es igual a la demanda total y que toda la cantidad del recurso disponible en un origen debe ser transportada a un único destino. El coste de transporte de una unidad de recurso depende del origen y el destino y viene dado por una función de coste c. El problema de transporte óptimo surge de encontrar el plan de transporte que minimice el coste total.

Se puede llevar una cabo una simplificación del problema suponiendo que la cantidad total de recurso que se debe transportar es 1 escalando las unidades. De esta forma se puede entender la disposición inicial del recurso como una distribución de probabilidad con soporte finito. De igual forma, la demanda del recurso en los destinos se puede ver como otra distribución de probabilidad con soporte finito. Entonces el problema planteado consiste en *transportar* una probabilidad en otra de forma que el coste del transporte sea mínimo en el sentido detallado anteriormente.

1.2. Formulación de Monge

En primer lugar se va a dar una definición que de sentido al concepto de *transportar* una probabilidad de un espacio $\mathcal{X}$ a otro $\mathcal{Y}$ por una aplicación $T : \mathcal{X} \to \mathcal{Y}$. Esta no es más que la ley inducida por T, la cual asigna medida a conjuntos de $\mathcal{Y}$.

Definición 1.2.1. Sea $\alpha \in P(\mathcal{X})$ y sea $T : \mathcal{X} \to \mathcal{Y}$ una aplicación medible. La ley inducida por T, que denotaremos por $T_{\#}\alpha$, está dada por

$$T_{\#}\alpha(A) = \alpha(T^{-1}(A)) = \alpha(\{x : T(x) \in A\}).$$

En los textos del ámbito del Transporte Óptimo es común emplear el término probabilidad *pushforward* para referirse a $T_{\#}\alpha$.

Es sencillo comprobar que $T_{\#}\alpha$ define una probabilidad en $\mathcal{Y}$ a partir de las propiedades de las imágenes inversas de una aplicación. Evidentemente $T_{\#}\alpha(A) \geq 0$ para cada A medible. Además, se satisface $T_{\#}\alpha(\mathcal{Y}) = \alpha(T^{-1}(\mathcal{Y})) = \alpha(\mathcal{X}) = 1$ y

$$T_{\#}\alpha\left(\sum_{i=1}^{\infty} A_n\right) = \alpha\left(T^{-1}\left(\sum_{i=1}^{\infty} A_n\right)\right) = \alpha\left(\sum_{i=1}^{\infty} T^{-1}\left(A_n\right)\right) = \sum_{i=1}^{\infty} \alpha(T^{-1}(A_n)) = \sum_{i=1}^{\infty} T_{\#}\alpha(A_n).$$

Con la notación que hemos introducido podemos formular el problema de transporte óptimo en le caso general sin restringirnos a probabilidades con soporte finito.

Definición 1.2.2 (Formulación de Monge del problema de transporte óptimo)**.** Sean $\alpha \in P(\mathcal{X})$, $\beta \in P(\mathcal{Y})$ y $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ una función de coste medible. El problema de transporte óptimo de Monge se formula como el siguiente problema de optimización

$$\inf_{T_{\#}\alpha=\beta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x, T(x))\, d\alpha(x) \right\} \tag{1.1}$$

donde $T : \mathcal{X} \to \mathcal{Y}$ es una aplicación medible.

Podemos ver que el problema definido en 1.2.2 abarca la situación que se ha planteado en la motivación inicial. Dado la situación descrita en la sección 1.2, se toma $\mathcal{X} = \{x_1, \ldots, x_n\}$ el conjunto de orígenes, $\mathcal{Y} = \{y_1, \ldots, y_m\}$ el conjunto de destinos, $\alpha = \sum_{i=1}^{n} a_n \delta_{x_n}$ la distribución inicial del recurso que se quiere transportar y $\beta = \sum_{i=1}^{m} b_m \delta_{y_m}$ la demanda del recurso que se quiere suplir. Se debe satisfacer

$$\sum_{i=1}^{n} a_n = 1 \; y \; \sum_{i=1}^{m} b_m = 1$$

para que α y β definan dos probabilidades en $\mathcal{X}$ e $\mathcal{Y}$ respectivamente. Se debe interpretar a_i como la proporción de recurso que se encuentra en el origen x_i y b_j se debe entender como la proporción de recurso que se demanda en el destino y_j. Consideramos la función de coste del transporte $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ dada por $c(x_i, y_j) = c_{i,j} \geq 0$, donde $c_{i,j}$ es el coste del transporte de una unidad de recurso desde el origen x_i hasta el destino y_j. Con esta notación el problema de transporte óptimo en la formulación de Monge se escribe

$$\inf_{T_{\#}\alpha=\beta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x, T(x))\, d\alpha(x) \right\} = \inf_{T_{\#}\alpha=\beta} \left\{ \sum_{i=1}^{n} c_{x_i, T(x_i)} \right\} \tag{1.2}$$

donde la restricción $T_{\#}\alpha = \beta$ es equivalente a $b_j = \beta(y_j) = \alpha(T^{-1}(y_j)) = \sum_{T(x_i)=y_j} a_i$.

La expresión (1.2) permite comprobar que la formulación obtenida coincide con la del problema inicial. En este sentido, la formulación de Monge generaliza el problema de transporte admitiendo que las distribuciones iniciales y finales tengan soporte infinito.

Observación 1. El problema de transporte óptimo en la formulación de Monge no es un problema de optimización convexa ya que en general la función objetivo no es convexa. Este hecho dificulta la resolución del problema que además puede no ser factible.

Podemos ver un ejemplo de esta situación si consideramos el siguiente problema de transporte: Tomamos un espacio unipuntual $\mathcal{X} = \{x\}$ y un espacio $\mathcal{Y} = \{y_1, y_2\}$ formado por dos puntos. Denotamos por α a la probabilidad sobre $\mathcal{X}$ que asigna masa 1 al punto x, y consideramos la probabilidad β que asigna masa $\frac{1}{2}$ a cada punto de $\mathcal{Y}$. Podemos notar que para cualquier aplicación T entre $\mathcal{X}$ e $\mathcal{Y}$ la ley inducida sobre $\mathcal{Y}$ asigna probabilidad 1 al punto y_1 o bien al punto y_2. Por lo tanto, se tiene $T_{\#}\alpha \neq \beta$ para cada T y en consecuencia el problema de transporte óptimo en la formulación de Monge no es factible.

1.3. Relajación de Kantorovich

En esta sección se va a introducir una formulación del problema del transporte óptimo que fue propuesta por Kantorovich. Este nuevo planteamiento surge como alternativa al que propuso Monge para evitar los problemas que hemos destacado en la Observación1. La formulación que vamos a desarrollar transforma el problema de transporte óptimo en un problema de optimización convexa lo que asegura la existencia de soluciones.

El éxito de la formulación de Kantorovich del problema de transporte óptimo radica en considerar planes de transporte óptimo en vez de aplicaciones de transporte óptimo. Es decir, si estudiamos el problema propuesto en la motivación inicial, podemos admitir que se transporten recursos a más de un destino desde un mismo origen. De esta forma, el conjunto en el cual debemos minimizar el coste es el conjunto de planes de transporte. Formalizamos este concepto en la siguiente definición.

Definición 1.3.1. Sean $\alpha \in \mathcal{P}(\mathcal{X}), \beta \in \mathcal{P}(\mathcal{Y})$. Consideramos las proyecciones $P_{\mathcal{X}}$ y $P_{\mathcal{Y}}$ sobre la primera y la segunda componente respectivamente. Denotamos por

$$\Pi(\alpha, \beta) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : (P_{\mathcal{X}})_{\#}\pi = \alpha \text{ y } (P_{\mathcal{Y}})_{\#}\pi = \beta\}$$

al conjunto de las probabilidades en el espacio producto $\mathcal{X} \times \mathcal{Y}$ con distribuciones marginales α y β respectivamente. Cada uno de sus elementos se denomina plan de transporte entre α y β.

Definición 1.3.2 (Relajación de Kantorovich)**.** Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades y sea $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ una función de coste medible. Se dice que

$$\mathcal{L}_c(\alpha, \beta) \overset{\text{def.}}{=} \inf_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\pi(x, y) \tag{1.3}$$

es el coste de transporte óptimo.

Se van a desarrollar una serie de resultados que van a permitir probar, bajo ciertas condiciones muy generales, que el inferior de la expresión (1.3) es en realidad un mínimo y que por lo tanto existe el plan de transporte óptimo. Recordamos que un espacio topológico $\mathcal{X}$ se dice que es polaco si existe una distancia d en $\mathcal{X}$ que induce la topología y $(\mathcal{X}, d)$ es un espacio métrico completo y separable. Un ejemplo de espacio polaco es $\mathbb{R}^n$. En los espacios polacos es válido el Teorema de Porkhorov el cual se va a emplear en la demostración de varios resultados de la sección. En el apéndice se enuncia este teorema junto con un resumen de conceptos y propiedades relativas a la convergencia débil.

Lema 1.3.1. Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos y sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. El conjunto de distribuciones conjuntas con marginales α y β, $\Pi(\alpha, \beta)$, es cerrado para la topología débil.

Demostración. Sea $\{\pi_k\}_{k=1}^{\infty}$ una sucesión de $\Pi(\alpha, \beta)$ que converge débilmente a π. Para cada $k \geq 1$ se tiene que

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x)\, d\pi_k(x, y) = \int_{\mathcal{X}} f(x)\, d\alpha(x)$$
$$\int_{\mathcal{X}\times\mathcal{Y}} g(y)\, d\pi_k(x, y) = \int_{\mathcal{Y}} g(y)\, d\beta(y)$$

para cada $f \in \mathcal{C}_b(\mathcal{X})$ y $g \in \mathcal{C}_b(\mathcal{Y})$. Tomando límites en estas expresiones podemos deducir que

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x)\, d\pi(x,y) = \lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} f(x)\, d\pi_k(x,y) = \int_{\mathcal{X}} f(x)\, d\alpha(x)$$
$$\int_{\mathcal{X}\times\mathcal{Y}} g(y)\, d\pi(x,y) = \lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} g(y)\, d\pi_k(x,y) = \int_{\mathcal{Y}} g(y)\, d\beta(y)$$

para cada $f \in \mathcal{C}_b(\mathcal{X})$ y $g \in \mathcal{C}_b(\mathcal{Y})$. Sea $A \subset \mathcal{X}$ un conjunto medible. Podemos tomar una sucesión creciente $\{f_k\}_{k=1}^{\infty}$ de funciones continuas y acotadas en $\mathcal{X}$ que converge puntualmente a la función indicatriz de A (ver la Proposición A.0.2). Entonces por el Teorema de la convergencia monótona se tiene que

$$\begin{aligned}\pi(A\times\mathcal{Y}) &= \int_{A\times\mathcal{Y}} d\pi(X,y) = \lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} f_k(x)\, d\pi(x,y)\\ &= \lim_{k\to\infty} \int_{\mathcal{X}} f_k(x)\, d\alpha(x) = \int_A d\alpha(x) = \alpha(A)\end{aligned}$$

De forma análoga se puede comprobar que $\pi(\mathcal{X}\times B) = \beta(B)$ para cada subconjunto $B \subset \mathcal{Y}$ medible. En consecuencia, $\pi \in \Pi(\alpha,\beta)$ y por lo tanto este conjunto es cerrado. □

Lema 1.3.2. Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos y sean $\mathcal{P} \subset \mathcal{P}(\mathcal{X})$ y $\mathcal{Q} \subset \mathcal{P}(\mathcal{Y})$ dos familias ajustadas. Entonces, el conjunto

$$\Pi(\mathcal{P},\mathcal{Q}) = \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : (P_{\mathcal{X}})_{\#}\pi \in \mathcal{P} \text{ y } (P_{\mathcal{Y}})_{\#}\pi \in \mathcal{Q}\}$$

de todas las distribuciones conjuntas con marginales en $\mathcal{P}$ y $\mathcal{Q}$ es ajustado en $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$.

Demostración. Fijamos un $\varepsilon > 0$. Por ser $\mathcal{P}$ y $\mathcal{Q}$ ajustados, existen subconjuntos compactos $K_\varepsilon \subset \mathcal{X}$ y $L_\varepsilon \subset \mathcal{Y}$ tales que

$$\mu(\mathcal{X}\setminus K_\varepsilon) \le \frac{\varepsilon}{2} \qquad \text{y} \qquad \nu(\mathcal{Y}\setminus L_\varepsilon) \le \frac{\varepsilon}{2}$$

para cada $\mu \in \mathcal{P}$ y $\nu \in \mathcal{Q}$. Sea $\pi \in \Pi(\mathcal{P},\mathcal{Q})$ una distribución con marginales $\mu \in \mathcal{P}$ y $\nu \in \mathcal{Q}$. Se tiene que

$$(\mathcal{X}\times\mathcal{Y})\setminus(K_\varepsilon\times L_\varepsilon) \subset ((\mathcal{X}\setminus K_\varepsilon)\times\mathcal{Y}) \cup (\mathcal{X}\times(\mathcal{Y}\setminus L_\varepsilon))$$

y por lo tanto

$$\pi((\mathcal{X}\times\mathcal{Y})\setminus(K_\varepsilon\times L_\varepsilon)) \le \pi((\mathcal{X}\setminus K_\varepsilon)\times\mathcal{Y}) + \pi(\mathcal{X}\times(\mathcal{Y}\setminus L_\varepsilon)) = \mu(\mathcal{X}\setminus K_\varepsilon) + \nu(\mathcal{Y}\setminus L_\varepsilon) < \varepsilon.$$

El conjunto $K_\varepsilon \times L_\varepsilon$ es compacto por ser producto de compactos con lo que se prueba que $\Pi(\mathcal{P},\mathcal{Q})$ es ajustado. □

Definición 1.3.3. Sean $\mathcal{X}$ un espacio polaco y $x_0 \in \mathcal{X}$. Sea $f : \mathcal{X} \to \mathbb{R}$ una función.

1. Se dice que f es superiormente semicontinua en x_0 si $\limsup_{x\to x_0} f(x) \le f(x_0)$.
2. Se dice que f es inferiormente semicontinua en x_0 si $\liminf_{x\to x_0} f(x) \ge f(x_0)$.

Se dice que f es es superiormente (inferiormente resp.) semicontinua si f es es superiormente (inferiormente resp.) semicontinua en x_0 para cada $x_0 \in \mathcal{X}$.

Lema 1.3.3. Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos. Sean $h : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{-\infty\}$ una función superiormente semicontinua y $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ una función de coste inferiormente semicontinua tal que $h \le c$. Sea $\{\pi_k\}_{k=1}^{\infty}$ una sucesión en $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ que converge débilmente a $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$. Si se verifica que

$$h \in \bigcap_{k=1}^{\infty} L^1(\pi_k) \cap L^1(\pi) \quad \text{y} \quad \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\, d\pi_k(x,y) \xrightarrow{w} \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\, d\pi(x,y),$$

entonces se satisface

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,d\pi(x,y) \leq \liminf_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,d\pi_k(x,y). \tag{1.4}$$

Demostración. La condición $h \leq c$ nos permite trabajar con $c-h$ la cual es una función no negativa interiormente semicontinua. Por el Teorema A.0.4 sabemos que existe una sucesión creciente $\{c_l\}_{l=1}^{\infty}$ de funciones continuas no negativas y acotadas que convergen puntualmente a $c-h$. Por el Teorema de la convergencia monótona y por definición de convergencia débil se tienen las siguientes igualdades:

$$\int_{\mathcal{X}\times\mathcal{Y}} (c-h)(x,y)\,d\pi(x,y) = \lim_{l\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_l(x,y)\,d\pi(x,y) = \lim_{l\to\infty} \left(\lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_l(x,y)\,d\pi_k(x,y) \right) \tag{1.5}$$

Además, para cada $l \geq 1$ se tiene que $c_l \leq c-h$, de lo que se deduce

$$\int_{\mathcal{X}\times\mathcal{Y}} c_l(x,y)\,d\pi_k(x,y) \leq \int_{\mathcal{X}\times\mathcal{Y}} (c-h)(x,y)\,d\pi_k(x,y).$$

Tomando límites en la desigualdad anterior se deduce que

$$\lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_l(x,y)\,d\pi_k(x,y) \leq \liminf_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} (c-h)(x,y)\,d\pi_k(x,y)$$

para cada $l \geq 1$ y en consecuencia tomando el límite cuando l tiende a infinito se obtiene

$$\lim_{l\to\infty} \left(\lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c_l(x,y)\,d\pi_k(x,y) \right) \leq \liminf_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} (c-h)(x,y)\,d\pi_k(x,y). \tag{1.6}$$

Las desigualdades (1.5) y (1.6) permiten deducir que

$$\begin{aligned} \int c\,d\pi - \int h\,d\pi = \int (c-h)\,d\pi &\leq \liminf_{k\to\infty} \int (c-h)\,d\pi_k = \liminf_{k\to\infty} \left(\int c\,d\pi_k - \int h\,d\pi_k \right) \\ &\leq \liminf_{k\to\infty} \left(\int c\,d\pi_k \right) - \lim_{k\to\infty} \left(\int h\,d\pi_k \right) \leq \liminf_{k\to\infty} \left(\int c\,d\pi_k \right) - \int h\,d\pi \end{aligned}$$

de donde se concluye el resultado. □

Los Lemas previos nos permiten probar el siguiente resultado el cual asegura la existencia de un plan de transporte óptimo, lo que destaca las ventajas de la formulación de Kantorovich frente a la formulación de Monge del problema de transporte óptimo.

Teorema 1.3.4 (Existencia del plan de transporte óptimo)**.** *Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos y sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. Sean $a : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ y $b : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ dos funciones superiormente semicontinuas tales que $a \in L^1(\alpha)$ y $b \in L^1(\beta)$. Sea $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ una función de coste inferiormente semicontinua tal que $a(x) + b(x) \leq c(x,y)$ para cada $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Entonces existe $\pi_0 \in \Pi(\alpha,\beta)$ que minimiza la expresión* (1.3).

Demostración. En primer lugar, vamos a probar que en las condiciones dadas existe el ínfimo definido en la expresión (1.3). Sea $\alpha\otimes\beta$ la medida producto de α y β. Podemos ver que $\alpha\otimes\beta \in \Pi(\alpha,\beta)$ ya que esta medida satisface la condición de las probabilidades marginales. Por lo tanto el conjunto

$$\left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,d\pi(x,y) \,:\, \pi \in \Pi(\alpha,\beta) \right\} \tag{1.7}$$

es no vacío. Para probar que existe el ínfimo del conjunto basta ver que este conjunto está acotado inferiormente. Dada $\pi \in \Pi(\alpha, \beta)$, se satisface

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) \geq \int_{\mathcal{X}\times\mathcal{Y}} a(x) + b(y)\, d\pi(x,y) = \int_{\mathcal{X}} a(x)\, d\alpha(x) + \int_{\mathcal{Y}} b(y)\, d\beta(y).$$

El término $\int_{\mathcal{X}} a(x)\, d\alpha(x) + \int_{\mathcal{Y}} b(y)\, d\beta(y)$ es un número real ya que $a \in L^1(\mathcal{X})$ y $b \in L^1(\mathcal{Y})$. En consecuencia existe el inferior de (1.7).

Para probar que el mínimo de (1.7) se alcanza, consideramos los subconjuntos $\{\alpha\} \subset \mathcal{P}(\mathcal{X})$ y $\{\beta\} \subset \mathcal{P}(\mathcal{Y})$. Por ser $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos, se tiene que los subconjuntos considerados son ajustados en $\mathcal{P}(\mathcal{X})$ y $\mathcal{P}(\mathcal{Y})$ respectivamente. Por el Lema 1.3.2 podemos deducir que $\Pi(\alpha, \beta)$ es ajustado en $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ de lo que deducimos que es precompacto en $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ por el Teorema de Prokhorov. Es decir, su adherencia en $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ es compacta para la topología débil. Por el Lema 1.3.1 deducimos que el conjunto de distribuciones conjuntas con marginales α y β coincide con su adherencia y por lo tanto es compacto para la topología débil.

Sea $\{\pi_k\}_{k=1}^{\infty}$ una sucesión de $\mathcal{P}(\alpha, \beta)$ para la cual $\int c\, d\pi_k$ converge al inferior de (1.7). Por la compacidad de $\mathcal{P}(\alpha, \beta)$ podemos extraer una subsucesión π_{k_n} de $\{\pi_k\}_{k=1}^{\infty}$ convergente débilmente a $\pi_0 \in \mathcal{P}(\alpha, \beta)$. Tomando $h(x, y) = a(x) + b(y)$ podemos comprobar que h es superiormente semicontinua y $h \leq c$. Además, se verifica que $h \in L^1(\pi_{k_n})$ y $h \in L^1(\pi_0)$ ya que

$$\int_{\mathcal{X}\times\mathcal{Y}} h\, d\pi_{k_n} = \int_{\mathcal{X}\times\mathcal{Y}} (a(x) + b(y))\, d\pi_{k_n} = \int_{\mathcal{X}} a\, d\alpha + \int_{\mathcal{Y}} b\, d\beta$$
$$\int_{\mathcal{X}\times\mathcal{Y}} h\, d\pi_0 = \int_{\mathcal{X}\times\mathcal{Y}} (a(x) + b(y))\, d\pi_0 = \int_{\mathcal{X}} a\, d\alpha + \int_{\mathcal{Y}} b\, d\beta$$

Entonces, por el Lema 1.3.3 se tiene que

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi_0(x,y) \leq \liminf_{n\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi_{k_n}(x,y) = \lim_{k\to\infty} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi_k(x,y) = \mathcal{L}_c(\alpha, \beta).$$

En consecuencia, el mínimo se alcanza en π_0. □

1.4. Distancia de Wasserstein

En esta sección se va a introducir una distancia entre probabilidades de un mismo espacio $\mathcal{X}$. Esta goza de buenas propiedades lo que la hacen ser una buena forma candidata para comparar distribuciones en contextos de aprendizaje automático y tratamiento de imágenes.

Definición 1.4.1. Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $p \geq 1$. Decimos que

$$\int_{\mathcal{X}} d(x, x_0)^p\, d\alpha(x) \tag{1.8}$$

es el momento de orden p de α respecto del punto $x_0 \in \mathcal{X}$. Decimos que una probabilidad α tiene momento de orden p finito si la integral (1.8) es finita para algún $x_0 \in \mathcal{X}$. Denotaremos por

$$\mathcal{P}_p(\mathcal{X}) = \{\alpha \in \mathcal{P}(\mathcal{X}) \ : \ \alpha \text{ tiene momento de orden } p \text{ finito}\}$$

al subconjunto de probabilidades en $\mathcal{X}$ con momento de orden p finito.

Proposición 1.4.1. *Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sea $p \geq 1$. Sea $\alpha \in \mathcal{P}_p(\mathcal{X})$. Entonces, para cada $y \in \mathcal{X}$ la integral*

$$\int_{\mathcal{X}} d(x, y)^p\, d\alpha(x)$$

es finita.

Demostración. Dado que α tiene momento de orden p finito, existe $x_0 \in \mathcal{X}$ tal que la integral $\int_{\mathcal{X}} d(x, x_0)^p \, d\alpha(x)$ es finita. Aplicando la desigualdad

$$d(x, y)^p \leq 2^{p-1}(d(x, x_0)^p + d(x_0, y)^p) \tag{1.9}$$

se concluye que

$$\begin{aligned}\int_{\mathcal{X}} d(x, y)^p \, d\alpha(x) &\leq \int_{\mathcal{X}} 2^{p-1}(d(x, x_0)^p + d(x_0, y)^p) \, d\alpha(x) \\ &= 2^{p-1} \int_{\mathcal{X}} d(x, x_0)^p \, d\alpha(x) + 2^{p-1} d(x_0, y)^p \int_{\mathcal{X}} d\alpha(x) < \infty.\end{aligned}$$

□

Observación 2. Fijado un $x_0 \in \mathcal{X}$ arbitrario basta comprobar si la integral (1.8) es finita para probar si $\alpha \in \mathcal{P}(\mathcal{X})$ tiene momento de orden p finito.

Incluimos un lema que nos va a ser útil para probar que la distancia de Wasserstein satisface la desigualdad triangular.

Lema 1.4.1 (Lema de pegado)**.** Sean $\mathcal{X}_1, \mathcal{X}_2$ y $\mathcal{X}_3$ tres espacios polacos y $\mu_i \in \mathcal{P}(\mathcal{X}_i)$ tres probabilidades. Consideramos las proyecciones sobre las dos primeras componentes y sobre las dos últimas componentes del espacio producto $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$:

$$P_{\mathcal{X}_1 \times \mathcal{X}_2}(x_1, x_2, x_3) = (x_1, x_2) \quad \text{y} \quad P_{\mathcal{X}_2 \times \mathcal{X}_3}(x_1, x_2, x_3) = (x_2, x_3).$$

Supongamos que se tienen $\pi_{1,2} \in \Pi(\mu_1, \mu_2)$ y $\pi_{2,3} \in \Pi(\mu_2, \mu_3)$. Entonces existe una distribución $\pi_{1,2,3} \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3)$ tal que

$$(P_{\mathcal{X}_1 \times \mathcal{X}_2})_{\#} \pi_{1,2,3} = \pi_{1,2} \quad \text{y} \quad (P_{\mathcal{X}_2 \times \mathcal{X}_3})_{\#} \pi_{1,2,3} = \pi_{2,3}.$$

Demostración. El argumento que se emplea en la demostración del lema se puede encontrar en [4, Chap.1]. □

Proposición 1.4.2. *Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sea $p \geq 1$. La expresión*

$$\mathcal{W}_p(\alpha, \beta) = \left(\inf_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\pi_(x, y) \right)^{\frac{1}{p}} \tag{1.10}$$

define una distancia en $\mathcal{P}_p(\mathcal{X})$.

Demostración. En primer lugar podemos ver que si tomamos como función de coste

$$c(x, y) = d(x, y)^p,$$

podemos escribir $\mathcal{W}_p(\alpha, \beta)^p = \mathcal{L}_c(\alpha, \beta)$. Dado que la distancia entre dos puntos es no negativa, se tiene que c es no negativa y por el Teorema 1.3.4 podemos concluir que existe una distribución conjunta π en $\mathcal{X} \times \mathcal{X}$ con marginales α y β tal que

$$\int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\pi = \mathcal{W}_p(\alpha, \beta)^p.$$

Veamos que $\mathcal{W}_p(\alpha, \beta)$ es finito si α, β tienen momentos de orden p finitos, es decir, si

$$\int_{\mathcal{X}} d(x, x_0)^p \, d\alpha(x) < \infty \quad \text{y} \quad \int_{\mathcal{X}} d(x, x_0)^p \, d\beta(x)$$

para un $x_0 \in \mathcal{X}$ arbitrario. Consideramos una distribución conjunta $\pi \in \Pi(\alpha, \beta)$. Aplicando la desigualdad (1.9) obtenemos

$$\begin{aligned}\mathcal{W}_p^p(\alpha,\beta) &\leq \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\pi(x,y) \leq \int_{\mathcal{X}} 2^{p-1}(d(x,x_0)^p + d(x_0,y)^p) \, d\pi(x,y) \\ &\leq 2^{p-1}\left(\int_{\mathcal{X}} d(x,x_0)^p \, d\alpha(x) + \int_{\mathcal{X}} d(x_0,y)^p \, d\beta(y)\right) < \infty.\end{aligned}$$

Solo falta probar que $\mathcal{W}_p$ define una distancia en $\mathcal{P}_p(\mathcal{X})$. Para ello consideramos α, β dos probabilidades en $\mathcal{X}$ con momentos de orden p finitos.

- Resulta sencillo comprobar que $\mathcal{W}_p(\alpha,\beta) \geq 0$ por ser la función de coste no negativa. Además, si $\mathcal{W}_p(\alpha,\beta) = 0$, consideramos $\pi \in \Pi(\alpha,\beta)$ en la que se alcanza el mínimo de (1.7). Se tiene que

  $$\mathcal{W}_p(\alpha,\beta)^p = \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\pi = 0$$

  de donde se deduce que $d(x,y)^p = 0$ π-c.s. Es decir, el conjunto $\{(x,y) \in \mathcal{X}\times\mathcal{X} \, : \, x \neq y\}$ tiene probabilidad nula. Por lo tanto, toda la masa de la probabilidad π está concentrada en la diagonal $\Delta = \{(x,x) \, : \, x \in \mathcal{X}\}$ de $\mathcal{X}\times\mathcal{X}$. Entonces para cada $A \subset \mathcal{X}$ medible se tiene

  $$\alpha(A) = \pi(A\times X) = \pi((A\times\mathcal{X})\cap\Delta) = \pi((\mathcal{X}\times A)\cap\Delta) = \pi(\mathcal{X}\times A) = \beta(A),$$

  de donde se deduce que $\alpha = \beta$.

- Consideramos la aplicación medible $T : \mathcal{X}\times\mathcal{X} \to \mathcal{X}\times\mathcal{X}$ dada por $T(x,y) = (y,x)$. Se tiene que la aplicación $T_\# : \Pi(\alpha,\beta) \to \Pi(\beta,\alpha)$ es biyectiva. Es fácil, comprobar este hecho notando que $T_\#(T_\#\pi) = \pi$.

  Además, dado $\pi \in \Pi(\alpha,\beta)$ se tiene que

  $$\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\pi(x,y) = \int_{\mathcal{X}\times\mathcal{X}} d(T(x,y))^p \, d\pi(x,y) = \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d(T_\#\pi)(x,y)$$

  puesto que $d\circ T = d$. De esta forma, el mínimo de (1.7) es el mismo cuando π recorre $\Pi(\alpha,\beta)$ y $\Pi(\beta,\alpha)$. En consecuencia, $\mathcal{W}_p$ es una función simétrica.

- Para probar la desigualdad triangular tomamos $\gamma \in \mathcal{P}_p(\mathcal{X})$. Consideramos las distribuciones conjuntas $\pi_{1,2}$, en la que se alcanza el mínimo de (1.7) en $\Pi(\alpha,\gamma)$, y $\pi_{2,3}$, en la que se alcanza el mínimo de (1.7) en $\Pi(\gamma,\beta)$. En estas condiciones y por ser $\mathcal{X}$ un espacio polaco, podemos aplicar el Lema de pegado (1.4.1) que nos asegura la existencia de

  $$\pi_{1,2,3} \in \mathcal{P}(\mathcal{X}\times\mathcal{X}\times\mathcal{X})$$

  que satisface

  $$(P_{\mathcal{X}_1\times\mathcal{X}_2})_\#\pi_{1,2,3} = \pi_{1,2} \quad \text{y} \quad (P_{\mathcal{X}_2\times\mathcal{X}_3})_\#\pi_{1,2,3} = \pi_{2,3}.$$

  Consideramos la distribución conjunta $\pi_{1,3} \in \mathcal{P}(\alpha,\beta)$ dada por

  $$\pi_{1,3} = (P_{\mathcal{X}_1\times\mathcal{X}_3})_\#\pi_{1,2,3}\,,$$

donde $P_{\mathcal{X}_1\times\mathcal{X}_3}$ denota la proyección sobre la primera y última componente. Entonces, aplicando la desigualdad de Minkowski se tiene que

$$\begin{aligned}
\mathcal{W}_p(\alpha,\beta) &\leq \left(\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\pi_{1,3}(x,y)\right)^{\frac{1}{p}} = \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\pi_{1,2,3}(x,z,y)\right)^{\frac{1}{p}} \\
&\leq \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} (d(x,z)+d(z,y))^p \, d\pi_{1,2,3}(x,z,y)\right)^{\frac{1}{p}} \\
&\leq \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} d(x,z)^p \, d\pi_{1,2,3}(x,z,y)\right)^{\frac{1}{p}} + \left(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} d(z,y)^p \, d\pi_{1,2,3}(x,z,y)\right)^{\frac{1}{p}} \\
&= \left(\int_{\mathcal{X}\times\mathcal{X}} d(x,z)^p \, d\pi_{1,2}(x,z)\right)^{\frac{1}{p}} + \left(\int_{\mathcal{X}\times\mathcal{X}} d(z,y)^p \, d\pi_{2,3}(z,y)\right)^{\frac{1}{p}} \\
&= \mathcal{W}_p(\alpha,\gamma) + \mathcal{W}_p(\gamma,\beta).
\end{aligned}$$

□

Definición 1.4.2. Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sea $p \geq 1$. Decimos que $\mathcal{W}_p$ es la distancia p de Wasserstein.

Observación 3. Podemos notar que si consideramos $\alpha, \beta, \gamma \in \mathcal{P}(\mathcal{X})$ arbitrarias sin que necesariamente tengan momentos de orden p finitos, se sigue satisfaciendo:

- $\mathcal{W}_p(\alpha,\beta) \geq 0$ y $\mathcal{W}_p(\alpha,\beta) = 0$ si y solo si $\alpha = \beta$
- $\mathcal{W}_p(\alpha,\beta) = \mathcal{W}_p(\beta,\alpha)$
- $\mathcal{W}_p(\alpha,\beta) \leq \mathcal{W}_p(\alpha,\gamma) + \mathcal{W}_p(\gamma,\beta)$

entendiendo estas expresiones en el contexto de $\mathbb{R} \cup \{\infty\}$. Esto se debe a que todos los argumentos que se han utilizado en la prueba de la Proposición 1.4.2 son válidos por ser la función de coste una función positiva.

La distancia de Wasserstein transporta la distancia d desde el $\mathcal{X}$ al espacio de distribuciones en $\mathcal{X}$, es decir, tiene en cuenta la localización de las probabilidades. Por ejemplo, si tomamos dos puntos $x_0, y_0 \in \mathcal{X}$ y consideramos las probabilidades δ_{x_0} y δ_{y_0} que concentran toda la masa en cada uno de estos puntos, podemos comprobar que la distancia p de Wasserstein entre ellas es

$$\mathcal{W}_p(\delta_{x_0}, \delta_{y_0}) = d(x_0, y_0) \tag{1.11}$$

para cualquier $p \geq 1$. Esta afirmación se comprueba fácilmente notando que $\Pi(\delta_{x_0}, \delta_{y_0}) = \{\delta_{(x_0,y_0)}\}$ y por lo tanto

$$\min_{\pi\in\Pi(\delta_{x_0},\delta_{y_0})} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\pi(x,y) = \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p \, d\delta_{(x_0,y_0)}(x,y) = d(x_0,y_0)^p.$$

Esta observación nos permite sumergir $\mathcal{X}$ en $\mathcal{P}_p(\mathcal{X})$ de forma isométrica asociando a cada punto $x \in \mathcal{X}$ la probabilidad δ_x. Esta propiedad dota a la distancia de Wasserstein de un sentido físico, el cual es una de las causas por las que es una buena medida para comparar distribuciones en aplicaciones como la modificación de imágenes.

En el caso particular en el que $\mathcal{X} = \mathbb{R}^n$, la distancia 2 de Wasserstein goza de buenas propiedades, una de ellas es que podemos obtener la distancia de Wasserstein entre dos probabilidades a partir de las probabilidades centradas.

Proposición 1.4.3. *Sean α, β dos probabilidades en $\mathbb{R}^n$ con momentos de orden 2 finitos. Denotamos por $\mathbf{m}_\alpha$ y $\mathbf{m}_\beta$ a los vectores de medias de α y β respectivamente y consideramos $\tilde{\alpha} = \alpha - \mathbf{m}_\alpha$ y $\tilde{\beta} = \beta - \mathbf{m}_\beta$ las probabilidades centradas. Entonces, se tiene que*

$$\mathcal{W}_2^2(\alpha, \beta) = \mathcal{W}_2^2(\tilde{\alpha}, \tilde{\beta}) + \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|_2^2. \tag{1.12}$$

Demostración. Consideramos la translación $\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)}$ de $\mathbb{R}^n \times \mathbb{R}^n$ dada por

$$\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y}) - (\mathbf{m}_\alpha, \mathbf{m}_\beta).$$

Podemos comprobar que la aplicación $\pi \to (\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)})_\# \pi$ es una biyección entre $\Pi(\alpha, \beta)$ y $\Pi(\tilde{\alpha}, \tilde{\beta})$: tomando $A, B \subset \mathbb{R}^n$ medibles podemos ver que

$$((\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)})_\# \pi)(A \times \mathbb{R}^n) = \pi(\tau_{(-\mathbf{m}_\alpha, -\mathbf{m}_\beta)}(A \times \mathbb{R}^n)) = \pi((A - \mathbf{m}_\alpha) \times \mathbb{R}^n) = \alpha(A - \mathbf{m}_\alpha) = \tilde{\alpha}(A),$$

$$((\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)})_\# \pi)(\mathbb{R}^n \times B) = \pi(\tau_{(-\mathbf{m}_\alpha, -\mathbf{m}_\beta)}(\mathbb{R}^n \times B)) = \pi(\mathbb{R}^n \times (B - \mathbf{m}_\beta)) = \beta(B - \mathbf{m}_\beta) = \tilde{\beta}(B).$$

Denotamos por $\tau_{(-\mathbf{m}_\alpha, -\mathbf{m}_\beta)}$ a la inversa de $\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)}$, la cual también es una translación. Además, se tiene que la aplicación $\pi \to (\tau_{(-\mathbf{m}_\alpha, -\mathbf{m}_\beta)})_\# \pi$ es la inversa de $(\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)})_\#(\cdot)$. En consecuencia, se tiene que

$$\begin{aligned}
\mathcal{W}_2^2(\tilde{\alpha}, \tilde{\beta}) &= \min_{\pi \in \Pi(\tilde{\alpha}, \tilde{\beta})} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 \, d\pi(x, y) \\
&= \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 \, d(\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)})_\# \pi(x, y).
\end{aligned}$$

Podemos desarrollar la última integral de la siguiente manera

$$\begin{aligned}
\int \|x - y\|_2^2 \, d(\tau_{(\mathbf{m}_\alpha, \mathbf{m}_\beta)})_\# \pi(x, y) &= \int \|(x - y) - (\mathbf{m}_\alpha - \mathbf{m}_\beta)\|_2^2 \, d\pi(x, y) \\
&= \int \|x - y\|_2^2 + \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|_2^2 - 2\langle x - y, \mathbf{m}_\alpha - \mathbf{m}_\beta \rangle \, d\pi(x, y).
\end{aligned}$$

Notando que $\int -2\langle x - y, \mathbf{m}_\alpha - \mathbf{m}_\beta \rangle \, d\pi(x, y) = -2\langle \int (x - y) \, d\pi(x, y), \mathbf{m}_\alpha - \mathbf{m}_\beta \rangle$ y que

$$\int (x - y) \, d\pi(x, y) = \int x \, d\alpha(x) - \int x \, d\beta(y) = \mathbf{m}_\alpha - \mathbf{m}_\beta$$

se deduce el resultado. □

La fórmula (1.12), que relaciona la distancia 2 de Wasserstein entre dos probabilidades y la distancia 2 de Wasserstein entre estas probabilidades centradas, pone en manifiesto la capacidad de esta distancia para detectar la diferencia de localización de dos probabilidades. Cuanto mayor sea la distancia entre los centros de gravedad de dos probabilidades en $\mathbb{R}^n$, mayor será la ditancia 2 de Wasserstein entre ellas.

A continuación se van a enunciar 3 resultados relativos a la distancia de Wasserstein. En los dos primeros, se caracteriza la convergencia en $\mathcal{P}_p(\mathcal{X})$ con la distancia $\mathcal{W}_p$ y se analiza la topología de este espacio métrico. En el tercer resultado, se relaciona la distancia 2 de Wasserstein con el problema de transporte óptimo en la formulación de Monge.

Teorema 1.4.2 (Convergencia en $\mathcal{P}_p(\mathcal{X})$)**.** *Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sean $\{\alpha_k\}_{k=1}^\infty$ una sucesión en $\mathcal{P}_p(\mathcal{X})$ y $\alpha \in \mathcal{P}_p(\mathcal{X})$ para $p \geq 1$. Consideramos un punto $x_0 \in \mathcal{X}$ arbitrario. Entonces, son equivalentes*

1. $\alpha_k \xrightarrow{w} \alpha$ y $\int_{\mathcal{X}} d(x, x_0)^p \, d\alpha_k(x) \longrightarrow \int_{\mathcal{X}} d(x, x_0)^p \, d\alpha(x)$

2. $\mathcal{W}_p(\alpha_k, \alpha) \longrightarrow 0$.

Demostración. La prueba del teorema se encuentra en [4, Thm.6.9] □

Observación 4. En las condiciones del Teorema 1.4.2 podemos ver que para cada $y \in \mathcal{X}$ se tiene la convergencia de los momentos de orden p

$$\int_{\mathcal{X}} d(x,y)^p \, d\alpha_k(x) \longrightarrow \int_{\mathcal{X}} d(x,y)^p \, d\alpha(x).$$

Teorema 1.4.3. *Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sea $p \geq 1$. Entonces $\mathcal{P}_p(\mathcal{X})$ es un espacio métrico completo y separable.*

Demostración. La prueba del teorema se encuentra en [4, Thm.6.18] □

Teorema 1.4.4 (Existencia de la aplicación de transporte óptimo)**.** *Sean α, β dos probabilidades en $\mathbb{R}^n$ tales que α no da probabilidad a subespacios de dimensión menor o igual a $n-1$. Entonces existe una única solución al problema de transporte óptimo* (1.3) *con función de coste*

$$c(x,y) = \|x - y\|_2^2$$

la cual es de la forma $\pi = (Id, T)_{\#}\alpha$ para una cierta aplicación $T : \mathbb{R}^n \to \mathbb{R}^n$. Esta aplicación es única y esta caracterizada por ser el gradiente de una función convexa φ, es decir, $T = \nabla\varphi$.

Demostración. La demostración del teorema se puede encontrar en [4, Thm.9.4]. También se puede encontrar una prueba alternativa del resultado en [5]. □

Observación 5. El Teorema 1.4.4 indica que la distancia 2 de Wasserstein entre dos probabilidades α, β que satisfagan las hipótesis del teorema se escribe

$$\mathcal{W}_2(\alpha, \beta) = \left(\int_{\mathbb{R}^n} \|x - T(x)\|_2^2 \, d\alpha(x) \right)^{\frac{1}{2}}$$

donde T es la aplicación de Monge.

El Teorema 1.4.4 proporciona un caso muy general en el que existe la aplicación de transporte óptimo. Se trata de un resultado de gran importancia ya que en numerosas aplicaciones del problema transporte óptimo se presenta la situación descrita en este teorema.

Observación 6. Un caso particular en el que se aplica este resultado se tiene tomando dos probabilidades $\alpha, \beta \in \mathcal{P}_2(\mathbb{R})$ tal que α tiene función de distribución continua. Por el Teorema 1.4.4 sabemos que existe la aplicación de transporte óptimo de Monge T entre α y β, ya que α no da masa a puntos. La caracterización que se ha dado de T en el teorema y la estructura de las funciones convexas en la recta real permiten deducir la expresión de la aplicación de transporte óptimo:

$$T = F_\beta^{-1} \circ F_\alpha \tag{1.13}$$

donde F_α es la función de distribución de α y F_β^{-1} es la función cuantil de β. Este resultado es consecuencia de [6, Thm.2.9]: basta considerar la función estrictamente convexa $h(x) = x^2$.

Según se ha comprobado a lo largo del capítulo, la formulación de Kantorovich es adecuada para el estudio del transporte óptimo. Además, se ha visto la equivalencia de esta formulación con la de Monge en una situación específica. Por este motivo, podría parecer conveniente emplear la distancia de Wasserstein para aplicaciones del área del aprendizaje automático como medida de similitud de dos distribuciones. En estos casos prácticos,

se trabaja con distribuciones de probabilidad discretas donde el número de puntos del soporte es finito, la cual es una restricción propia del empleo de métodos numéricos. En el Capítulo 3, se estudiará el cálculo del coste óptimo en el caso discreto en la formulación de Kantorovich, el cual se corresponde con la resolución de un problema de programación lineal. Aquí surge el principal problema del enfoque de Kantorovich: los algoritmos convencionales para la obtención de la solución óptima de un problema de programación lineal no son lo suficientemente rápidos y no se pueden emplear de forma efectiva cuando el tamaño del problema es muy grande. Esta es la razón principal que motiva la introducción de la regularización entrópica, la cual se va a estudiar en el Capítulo 2 y sobre la cual se desarrollará un análisis del caso discreto en el Capítulo 3.

Capítulo 2

Regularización entrópica

Este capítulo está dedicado al estudio del problema de transporte entrópico, el cual surge de la introducción de una regularización o penalización entrópica al problema de transporte óptimo de Kantorovich. El factor de regularización va a permitir controlar la penalización introducida pudiendo recuperar el problema de transporte óptimo original disminuyendo el factor de regularización. Se va a estudiar las propiedades de esta formulación que posteriormente serán relevantes en el caso discreto, el cual se analizará en el Capítulo 3. La referencia que se ha empleado como base de este capítulo es [7].

Definición 2.0.1. Sean P y R dos probabilidades sobre un espacio medible $\mathcal{X}$. Si P es absolutamente continua respecto de R (ver Apéndice A), se define la divergencia de Kullback-Leibler

$$\mathrm{KL}\,(P|R) = \int_{\mathcal{X}} \log\left(\frac{dP}{dR}(x)\right)\, dP(x) \tag{2.1}$$

donde $\frac{dP}{dR}(x)$ es la derivada de Radon-Nikodym de P respecto de R (ver Apéndice A). Si P no es absolutamente continua respecto de R, se define $\mathrm{KL}\,(P|R) = \infty$.

Proposición 2.0.1. *Sean P y R probabilidades sobre un espacio medible $\mathcal{X}$. Entonces*

$$KL(P|R) \geq 0$$

y la igualdad se da si $P = R$.

Demostración. En primer lugar, vamos a probar que la divergencia de Kullback-Leibler es positiva. Podemos suponer que P es absolutamente continua respecto de R porque en caso contrario $\mathrm{KL}\,(P|R) = \infty$ por definición. Consideramos la función convexa $h(z) = z\log(z)$ para $z > 0$. Observamos que se satisface la siguiente igualdad:

$$\int_{\mathcal{X}} h\left(\frac{dP}{dR}(x)\right)\, dR(x) = \int_{\mathcal{X}} \left(\frac{dP}{dR}(x)\right) \cdot \log\left(\frac{dP}{dR}(x)\right)\, dR(x) = \int_{\mathcal{X}} \log\left(\frac{dP}{dR}(x)\right)\, dP(x).$$

Además, se tiene que $\int_{\mathcal{X}} \frac{dP}{dR}(x)\, dR(x) = \int_{\mathcal{X}} dP(x) = 1$, de lo que se deduce que $h\left(\int_{\mathcal{X}} \frac{dP}{dR}(x)\, dR(x)\right) = 0$. Por ser h una función convexa se deduce de la desigualdad de Jensen que

$$\mathrm{KL}\,(P|R) = \int_{\mathcal{X}} h\left(\frac{dP}{dR}(x)\right)\, dR(x) \geq h\left(\int_{\mathcal{X}} \frac{dP}{dR}(x)\, dR(x)\right) = 0.$$

Dado que h es estrictamente convexa, se da la igualdad si y solo si $\frac{dP}{dR}(x)$ es constante. Es decir, se da la igualdad si y solo si $P = R$. □

Proposición 2.0.2. *Sean P, Q y R probabilidades sobre un espacio medible $\mathcal{X}$ tales que* $\log(\frac{dQ}{dR}(x)) \in L^1(P)$ *y $Q \sim R$. Entonces*

$$KL\,(P|R) = KL\,(P|Q) + \int_{\mathcal{X}} \log\left(\frac{dQ}{dR}(x)\right)\, dP(x). \tag{2.2}$$

Demostración. Debemos entender la expresión (2.2) como una igualdad formal, en la que si el lado izquierdo de la igualdad es ∞ entonces el lado derecho de la igualdad es ∞. Esta convención tiene sentido porque los términos $\mathrm{KL}\,(P|R)$, $\mathrm{KL}\,(P|Q)$ son positivos y la integral de la derecha siempre es finita.

Supongamos que P no es absolutamente continua respecto de R. En esta situación se tiene que $\mathrm{KL}\,(P|R) = \infty$. Dado que Q y R son equivalentes, se deduce que P no es absolutamente continua respecto de Q y por lo tanto $\mathrm{KL}\,(P|Q) = \infty$.

Supongamos que P es absolutamente continua respecto de R, entonces P es absolutamente continua respecto de Q por ser Q y R equivalentes. La entropía de P respecto de R se puede escribir

$$\begin{aligned}
\mathrm{KL}\,(P|R) &= \int_{\mathcal{X}} \log\left(\frac{dP}{dR}(x)\right)\, dP(x) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}(x)\cdot\frac{dQ}{dR}(x)\right)\, dP(x) \\
&= \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}(x)\right)\, dP + \int_{\mathcal{X}} \log\left(\frac{dQ}{dR}(x)\right)\, dP(x) \\
&= \mathrm{KL}\,(P|Q) + \int_{\mathcal{X}} \log\left(\frac{dQ}{dR}(x)\right)\, dP(x)
\end{aligned}$$

□

A continuación se va a presentar el problema de transporte entrópico o problema de transporte óptimo regularizado, que surge de añadir una penalización a la formulación de Kantorovich. Este término que se añade a la formulación (1.3) está dado por una divergencia de Kullback-Leibler.

Definición 2.0.2 (Regularización entrópica)**.** Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. Sean $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ una función de coste y $\varepsilon > 0$.

$$\mathcal{L}_c^{\varepsilon}(\alpha,\beta) \overset{\text{def.}}{=} \inf_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) + \varepsilon\,\mathrm{KL}\,(\pi|\alpha\otimes\beta) \tag{2.3}$$

Definición 2.0.3. Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. Sea $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ una función de coste acotada inferiormente y $\varepsilon > 0$. Denotamos por $\mathcal{K}_\varepsilon$ a la probabilidad del espacio producto dada por

$$\frac{d\mathcal{K}_\varepsilon}{d(\alpha\otimes\beta)}(x,y) = \lambda\, e^{-\frac{c(x,y)}{\varepsilon}}$$

donde $\lambda^{-1} = \int e^{-\frac{c}{\varepsilon}}\, d(\alpha\otimes\beta)$.

Proposición 2.0.3. *Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades y sea $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ una función de coste acotada inferiormente. El problema de transporte óptimo entrópico* (2.3) *se puede reformular como*

$$\mathcal{L}_c^{\varepsilon}(\alpha,\beta) = \inf_{\pi\in\Pi(\alpha,\beta)} \varepsilon\, KL\,(\pi|\mathcal{K}_\varepsilon) + \varepsilon\log(\lambda). \tag{2.4}$$

Demostración. En virtud de la Proposición 2.0.2 podemos escribir

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Y}} c\,d\pi + \varepsilon\,\mathrm{KL}(\pi|\alpha\otimes\beta) &= \int_{\mathcal{X}\times\mathcal{Y}} c\,d\pi + \varepsilon\,\mathrm{KL}(\pi|\mathcal{K}_\varepsilon) + \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\mathcal{K}_\varepsilon}{d(\alpha\otimes\beta)}\right) d\pi \\ &= \int_{\mathcal{X}\times\mathcal{Y}} c\,d\pi + \varepsilon\,\mathrm{KL}(\pi|\mathcal{K}_\varepsilon) + \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\lambda e^{-\frac{c}{\varepsilon}}\right) d\pi \\ &= \int_{\mathcal{X}\times\mathcal{Y}} c\,d\pi + \varepsilon\,\mathrm{KL}\left(\pi|\mathcal{K}_\varepsilon\right) - \int_{\mathcal{X}\times\mathcal{Y}} c\,d\pi + \varepsilon\log(\lambda) \\ &= \varepsilon\,\mathrm{KL}(\pi|\mathcal{K}_\varepsilon) + \varepsilon\log(\lambda)\end{aligned}$$

para una probabilidad $\pi \in \Pi(\alpha,\beta)$ que verifique $c \in L^1(\pi)$, de donde se deduce la equivalencia. □

Observación 7. La Proposición 2.0.3 nos proporciona una vía para probar que en (2.3) se alcanza el mínimo bajo condiciones muy generales. Además, a partir de esta representación y de la convexidad estricta del funcional $\mathrm{KL}(\,\cdot\,|\mathcal{K}_\varepsilon)$, se va a poder probar que este mínimo es único. Realmente se va a probar que

$$\mathrm{KL}(\,\cdot\,|\mathcal{K}_\varepsilon) \tag{2.5}$$

alcanza un único mínimo en $\Pi(\alpha,\beta)$, que es equivalente al resultado enunciado. Podemos observar que la probabilidad que minimiza (2.5) también minimiza (2.3).

Lema 2.0.1. Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. El conjunto de distribuciones conjuntas con marginales α y β, $\Pi(\alpha,\beta)$, es convexo.

Demostración. Sean $\mu,\nu \in \Pi(\alpha,\beta)$ y sea $\lambda \in [0,1]$. Recordamos que si definimos

$$((1-\lambda)\mu + \lambda\nu)(M) = (1-\lambda)\mu(M) + \lambda\nu(M)$$

para cada $M \subset \mathcal{X}\times\mathcal{Y}$ medible, entonces $((1-\lambda)\mu+\lambda\nu)$ define una probabilidad en $\mathcal{X}\times\mathcal{Y}$. Dados $A\subset\mathcal{X}$ y $B\subset\mathcal{Y}$ medibles se tiene que

$$((1-\lambda)\mu+\lambda\nu)(A\times\mathcal{Y}) = (1-\lambda)\mu(A\times\mathcal{Y}) + \lambda\nu(A\times\mathcal{Y}) = (1-\lambda)\alpha(A) + \lambda\alpha(A) = \alpha(A)$$

$$((1-\lambda)\mu+\lambda\nu)(\mathcal{X}\times B) = (1-\lambda)\mu(\mathcal{X}\times B) + \lambda\nu(\mathcal{X}\times B) = (1-\lambda)\beta(B) + \lambda\beta(B) = \beta(B)$$

de donde se se concluye que $((1-\lambda)\mu+\lambda\nu)$ pertenece a $\Pi(\alpha,\beta)$. □

A continuación, se va a introducir la distancia en variación total, la cual es una distancia en el conjunto de probabilidades de un espacio medible $\mathcal{X}$. También enunciaremos la desigualdad de Pinsker ya que los emplearemos en los lemas posteriores. Esta desigualdad relaciona la distancia en variación total entre dos probabilidades con su entropía relativa.

Definición 2.0.4. Sean P,Q dos probabilidades sobre un espacio medible $\mathcal{X}$. La expresión

$$\|P-Q\|_{TV} = \frac{1}{2}\cdot \sup_{\substack{M\subset\mathcal{X}\\ \text{medible}}} |P(M)-Q(M)|$$

define una distancia en $\mathcal{P}(\mathcal{X})$ acotada por 1 la cual se denomina distancia en variación total.

Observación 8. La convergencia en variación total implica la convergencia débil en espacios polacos donde se satisface el Teorema Portmanteau (Teorema A.0.1): es sencillo comprobar que si $\lim_{n\to\infty}\|P_n - P\|_{TV} = 0$, entonces $\lim_{n\to\infty} P_n(A) = P(A)$ y por lo tanto se verifica la condición 6 del Teorema Portmanteau.

Proposición 2.0.4 (Desigualdad de Pinsker)**.** *Sean P, Q dos probabilidades sobre un espacio medible $\mathcal{X}$. La distancia en variación total entre P y Q se acota por*

$$\|P - Q\|_{TV} \leq \sqrt{\frac{1}{2} KL\,(P|Q)}.$$

Demostración. Se puede encontrar una prueba de la desigualdad en [8]. □

Lema 2.0.2. Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. El conjunto de distribuciones conjuntas con marginales α y β, $\Pi(\alpha, \beta)$, es cerrado para la topología de la convergencia en variación total.

Demostración. Sea $\{\pi_n\}_{n=1}^{\infty}$ una sucesión en $\Pi(\alpha, \beta)$ que converge en variación total a π. Esto es

$$\|\pi_n - \pi\|_{TV} = \frac{1}{2} \cdot \sup_{M \text{ medible}} |\pi_n(M) - \pi(M)| \xrightarrow[n\to\infty]{} 0.$$

En particular para cada $A \subset \mathcal{X}$ y $B \subset \mathcal{Y}$ medibles se tiene que

$$|\pi_n(A \times \mathcal{Y}) - \pi(A \times \mathcal{Y})| = |\alpha(A) - \pi(A \times \mathcal{Y})| \xrightarrow[n\to\infty]{} 0$$
$$|\pi_n(\mathcal{X} \times B) - \pi(\mathcal{X} \times B)| = |\beta(B) - \pi(\mathcal{X} \times B)| \xrightarrow[n\to\infty]{} 0$$

de donde se deduce que $\pi(A \times \mathcal{Y}) = \alpha(A)$ y $\pi(\mathcal{X} \times B) = \beta(B)$. Concluimos que $\pi \in \Pi(\alpha, \beta)$ y por lo tanto es cerrado. □

Proposición 2.0.5. *Sea R una probabilidad sobre $\mathcal{X}$. El funcional*

$$KL\,(\,\cdot\,|R)$$

es inferiormente semicontinuo en el subespacio en el que es finito respecto a la convergencia débil.

Demostración. La prueba de este resultado se encuentra en [9, Thm.1]. □

Lema 2.0.3. Sea R una probabilidad sobre $\mathcal{X}$. El funcional

$$\mathrm{KL}\,(\,\cdot\,|R)$$

es estrictamente convexo en el subespacio en el que es finito. Es decir, si $\mathrm{KL}\,(P|R) < \infty$ y $\mathrm{KL}\,(Q|R) < \infty$ y $\lambda \in (0, 1)$, entonces

$$\mathrm{KL}\,((1-\lambda)P + \lambda Q|R) < (1-\lambda)\,\mathrm{KL}\,(P|R) + \lambda\,\mathrm{KL}\,(Q|R)\,.$$

Demostración. Consideramos la función $h(z) = z\log(z)$ la cual es estrictamente convexa ya que la segunda derivada es estrictamente positiva:

$$h''(z) = \frac{1}{z} > 0 \quad \forall z > 0.$$

Sean P y Q dos probabilidades tales que $\mathrm{KL}\,(P|R) < \infty$ y $\mathrm{KL}\,(Q|R) < \infty$ y sea $\lambda \in (0, 1)$. Podemos deducir que existen las derivadas de Radon-Nikodym $\frac{dP}{dR}$ y $\frac{dQ}{dR}$. Por ser h estrictamente convexa se tiene que

$$h\left(\frac{(1-\lambda)\,dP + \lambda\,dQ}{dR}\right) = h\left((1-\lambda)\frac{dP}{dR} + \lambda\frac{dQ}{dR}\right) < (1-\lambda)\,h\left(\frac{dP}{dR}\right) + \lambda\,h\left(\frac{dQ}{dR}\right),$$

de donde se deduce que

$$\begin{aligned}\mathrm{KL}\left((1-\lambda)P+\lambda Q|R\right) &= \int_{\mathcal{X}} h\left((1-\lambda)\frac{dP}{dR}+\lambda\frac{dQ}{dR}\right)dR \\ &< (1-\lambda)\int_{\mathcal{X}} h\left(\frac{dP}{dR}\right)dR+\lambda\int_{\mathcal{X}} h\left(\frac{dQ}{dR}\right)dR \\ &= (1-\lambda)\,\mathrm{KL}\left(P|R\right)+\lambda\,\mathrm{KL}\left(Q|R\right).\end{aligned}$$

□

Lema 2.0.4. Sea $\{P_n\}_{n=1}^{\infty}$ una sucesión de probabilidades en $\mathcal{X}$ tal que el límite

$$\lim_{n\to\infty}\mathrm{KL}(P_n|R)=L$$

existe y es finito. Sea $P_{m,n}=\frac{1}{2}(P_m+P_n)$ para cada par (m,n). Si $\liminf_{n,m\to\infty}\mathrm{KL}(P_{m,n}|R)\geq L$, entonces la sucesión $\{P_n\}_{n=1}^{\infty}$ converge en variación total.

Demostración. En primer lugar, notamos que por ser $\mathrm{KL}\left(\,\cdot\,|R\right)$ convexo se tiene

$$\mathrm{KL}(P_{m,n}|R)\leq\tfrac{1}{2}\mathrm{KL}\left(P_m|R\right)+\tfrac{1}{2}\mathrm{KL}\left(P_n|R\right)$$

de lo que tomando límites superiores se deduce que $\limsup_{m,n\to\infty}\mathrm{KL}(P_{m,n}|R)\leq L$. Entonces se debe satisfacer

$$\lim_{m,n\to\infty}\mathrm{KL}(P_{m,n}|R)=L.$$

Podemos observar que $P_{m,n}$ es absolutamente continua respecto de R para m y n suficientemente grandes y se tiene

$$\tfrac{dP_{m,n}}{dR}(x)=\tfrac{1}{2}\left(\tfrac{P_m}{dR}(x)+\tfrac{P_n}{dR}(x)\right),$$

lo que nos permite desarrollar las siguientes divergencias

$$\begin{aligned}\mathrm{KL}\left(P_m|R\right) &= \mathrm{KL}\left(P_m|P_{m,n}\right)+\int_{\mathcal{X}}\log\left(\frac{1}{2}\left(\frac{dP_m}{dR}+\frac{dP_n}{dR}\right)\right)\frac{dP_m}{dR}\,dR, \\ \mathrm{KL}\left(P_n|R\right) &= \mathrm{KL}\left(P_n|P_{m,n}\right)+\int_{\mathcal{X}}\log\left(\frac{1}{2}\left(\frac{dP_m}{dR}+\frac{dP_n}{dR}\right)\right)\frac{dP_n}{dR}\,dR.\end{aligned}$$

Sumando las dos expresiones anteriores obtenemos la igualdad

$$\begin{aligned}\mathrm{KL}\left(P_m|R\right)+\mathrm{KL}\left(P_n|R\right) &= \mathrm{KL}\left(P_m|P_{m,n}\right)+\mathrm{KL}\left(P_n|P_{m,n}\right)+ \\ &\quad+2\int_{\mathcal{X}}\log\left(\frac{1}{2}\left(\frac{dP_m}{dR}+\frac{dP_n}{dR}\right)\right)\frac{1}{2}\left(\frac{dP_m}{dR}+\frac{dP_n}{dR}\right)dR \\ &= \mathrm{KL}\left(P_m|P_{m,n}\right)+\mathrm{KL}\left(P_n|P_{m,n}\right)+2\mathrm{KL}\left(P_{m,n}|R\right).\end{aligned}$$

Tomando límites superiores en la expresión anterior obtenemos

$$2L=2L+\limsup_{m,n\to\infty}\mathrm{KL}\left(P_m|P_{m,n}\right)+\limsup_{m,n\to\infty}\mathrm{KL}\left(P_m|P_{m,n}\right)$$

de lo que se deduce que $\lim_{m,n\to\infty}\mathrm{KL}\left(P_m|P_{m,n}\right)=\lim_{m,n\to\infty}\mathrm{KL}\left(P_n|P_{m,n}\right)=0$. Por la desigualdad de Pinsker deducimos que

$$\begin{aligned}\|P_m-P_{m,n}\|_{TV} &\xrightarrow[m,n\to\infty]{} 0 \\ \|P_n-P_{m,n}\|_{TV} &\xrightarrow[m,n\to\infty]{} 0\end{aligned}$$

Por la desigualdad triangular concluimos

$$\|P_m - P_n\|_{TV} \leq \|P_m - P_{m,n}\|_{TV} + \|P_{m,n} - P_n\|_{TV} \xrightarrow[m,n\to\infty]{} 0.$$

Luego hemos probado que la sucesión $\{P_n\}_{n=1}^{\infty}$ es de Cauchy y por lo tanto convergente ya que $\mathcal{P}(\mathcal{X})$ es completo para la convergencia en variación total. La prueba de este resultado se encuentra en la Proposición A.0.1 del Apéndice A. □

Teorema 2.0.5 (Existencia y unicidad del transporte entrópico)**.** *Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos y sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades. Sean $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ una función de coste acotada inferiormente tal que $c \in L^1(\alpha \otimes \beta)$ y $\varepsilon > 0$. Entonces existe una única probabilidad $\pi_0 \in \Pi(\alpha, \beta)$ que minimiza la expresión* (2.3)*.*

Demostración. En los Lemas 2.0.1 y 2.0.2 se ha probado que $\mathcal{Q} = \Pi(\alpha, \beta)$ es convexo y cerrado para la convergencia en variación total. Además, podemos ver que

$$\mathrm{KL}(\alpha \otimes \beta | \mathcal{K}_\varepsilon) = \int_{\mathcal{X}\times\mathcal{Y}} \log(\lambda^{-1} e^{\frac{c}{\varepsilon}})\, d(\alpha \otimes \beta) = \int_{\mathcal{X}\times\mathcal{Y}} c\, d(\alpha \otimes \beta) - \log(\lambda) < \infty.$$

Consideramos ahora $\{\pi_n\}_{n=1}^{\infty}$ una sucesión para la cual $\mathrm{KL}(\pi_n|\mathcal{K}_\varepsilon)$ converge a

$$L = \inf_{\pi \in \Pi(\alpha,\beta)} \mathrm{KL}(\pi|\mathcal{K}_\varepsilon).$$

En primer lugar, notamos que $L \in \mathbb{R}_{\geq 0}$ porque debe ser menor que $\mathrm{KL}(\alpha \otimes \beta|\mathcal{K}_\varepsilon)$ y además el funcional $\mathrm{KL}(\cdot|\mathcal{K}_\varepsilon)$ es no negativo. Tomando ahora $\pi_{m,n} = \frac{1}{2}(\pi_m + \pi_n) \in \Pi(\alpha, \beta)$ podemos ver que $\mathrm{KL}(\pi_{m,n}|\mathcal{K}_\varepsilon) \geq L$ para cada $m \geq 1$ y $n \geq 1$ por ser L el inferior. Por lo tanto

$$\liminf_{m,n\to\infty} \mathrm{KL}(\pi_{m,n}|\mathcal{K}_\varepsilon) \geq L$$

y por el Lema 2.0.4 deducimos que $\{\pi_n\}_{n=1}^{\infty}$ converge en variación total a un cierto π_0 que debe pertenecer a $\Pi(\alpha, \beta)$ por ser cerrado. El Lema 2.0.5 y la Observación 8 implican que $\mathrm{KL}(\,\cdot\,|\mathcal{K}_\varepsilon)$ es inferiormente semicontinuo para la convergencia en variación total. Entonces, se tiene

$$\mathrm{KL}(\pi_0|\mathcal{K}_\varepsilon) \leq \liminf_{n\to\infty} \mathrm{KL}(\pi_n|\mathcal{K}_\varepsilon) = L$$

y por lo tanto el mínimo se alcanza en π_0.

La unicidad del mínimo se deduce de la convexidad estricta del $\mathrm{KL}(\cdot|\mathcal{K}_\varepsilon)$: si existiesen dos mínimos π_0 y π_1 distintos, entonces

$$\mathrm{KL}\left(\frac{1}{2}(\pi_0 + \pi_1)\middle|\mathcal{K}_\varepsilon\right) < \frac{1}{2}\mathrm{KL}(\pi_0|\mathcal{K}_\varepsilon) + \frac{1}{2}\mathrm{KL}(\pi_1|\mathcal{K}_\varepsilon) = L$$

lo cual es absurdo ya que $\frac{1}{2}(\pi_0 + \pi_1) \in \Pi(\alpha, \beta)$ y contradice la optimalidad de π_0 y π_1.

□

Proposición 2.0.6 (Convexidad estricta)**.** *Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios polacos y sean $\alpha_1, \alpha_2 \in \mathcal{P}(\mathcal{X})$ y $\beta_1, \beta_2 \in \mathcal{P}(\mathcal{Y})$ tales que no se dé simultáneamente*

$$\alpha_1 = \alpha_2 \quad y \quad \beta_1 = \beta_2. \tag{2.6}$$

Dado $\lambda \in (0, 1)$ consideramos las probabilidades $\mu = \lambda\alpha_1 + (1-\lambda)\alpha_2$ y $\nu = \lambda\beta_1 + (1-\lambda)\beta_2$ definidas en $\mathcal{X}$ e $\mathcal{Y}$ respectivamente. Sea $\varepsilon > 0$ un factor de regularización y $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ una función de coste acotada inferiormente tal que $c \in L^1(\alpha_1 \otimes \beta_1) \cap L^1(\alpha_2 \otimes \beta_2)$. Entonces

$$\mathcal{L}_c^\varepsilon(\mu, \nu) < \lambda\mathcal{L}_c^\varepsilon(\alpha_1, \beta_1) + (1-\lambda)\mathcal{L}_c^\varepsilon(\alpha_2, \beta_2)$$

Demostración. Sean $\pi_1 \in \Pi(\alpha_1, \beta_1)$ y $\pi_2 \in \Pi(\alpha_2, \beta_2)$ las distribuciones en las que se alcanzan los mínimos $\mathcal{L}_c^\varepsilon(\alpha_1, \beta_1)$ y $\mathcal{L}_c^\varepsilon(\alpha_2, \beta_2)$ respectivamente. Dado que no se verifica (2.6) por hipótesis, se tiene $\pi_1 \neq \pi_2$. Podemos ver que la probabilidad $\lambda\pi_1 + (1-\lambda)\pi_2$ pertenece a $\Pi(\mu, \nu)$:

$$(\lambda\pi_1 + (1-\lambda)\pi_2)(A \times \mathcal{Y}) = \lambda\pi_1(A \times \mathcal{Y}) + (1-\lambda)\pi_2(A \times \mathcal{Y}) = \lambda\alpha_1(A) + (1-\lambda)\alpha_2(A) = \mu(A)$$
$$(\lambda\pi_1 + (1-\lambda)\pi_2)(\mathcal{X} \times B) = \lambda\pi_1(\mathcal{X} \times B) + (1-\lambda)\pi_2(\mathcal{X} \times B) = \lambda\beta_1(B) + (1-\lambda)\beta_2(B) = \nu(B).$$

Además, por la convexidad estricta de $\mathrm{KL}\,(\,\cdot\,|\mathcal{K}_\varepsilon)$ se deduce

$$\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}\,(\pi|\mathcal{K}_\varepsilon) \leq \mathrm{KL}\,(\lambda\pi_1 + (1-\lambda)\pi_2|\mathcal{K}_\varepsilon) < \lambda\mathrm{KL}\,(\pi_1|\mathcal{K}_\varepsilon) + (1-\lambda)\mathrm{KL}\,(\pi_2|\mathcal{K}_\varepsilon)\,.$$

El resultado se deduce como consecuencia de la relación dada en la Proposición 2.0.3 y de que en π_i se alcanza el mínimo de $\mathrm{KL}\,(\,\cdot\,|\mathcal{K}_\varepsilon)$ en $\Pi(\alpha_i, \beta_i)$ para $i = 1, 2$. □

Teorema 2.0.6. *Sea $\{\varepsilon_l\}_{l=1}^\infty$ una sucesión de números estrictamente positivos que converge a 0. Sea π_l la única solución del problema de transporte entrópico con factor de regularización ε_l. Sea c una función de coste continua que verifique $c(x, y) \leq a(x) + b(y)$ para funciones $a \in L^1(\alpha)$ y $b \in L^1(\beta)$. Entonces el coste de transporte entrópico converge al coste de transporte óptimo, y si las soluciones del problema de transporte entrópico π_l convergen débilmente, lo hacen a una solución del problema de transporte óptimo. Además, en caso de que solo exista una única solución π^* del problema de transporte óptimo, se tiene $\pi_l \xrightarrow{w} \pi^*$.*

Demostración. Ver [7, Thm.5.10]. □

Definición 2.0.5. Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable y sea $p \geq 1$. Consideramos la función de coste $c(x, y) = d(x, y)^p$. Fijado el factor de regularización $\varepsilon > 0$, decimos que

$$\mathcal{W}_p^\varepsilon(\alpha, \beta) = \mathcal{L}_c^\varepsilon(\alpha, \beta)^{\frac{1}{p}} \tag{2.7}$$

es la distancia p de Wasserstein regularizada entre las probabilidades α y β.

Observación 9. Al contrario que la distancia p de Wasserstein, $\mathcal{W}_p^\varepsilon$ no es una distancia en $\mathcal{P}_p(\mathcal{X})$. Sin embargo, tomando ε suficientemente pequeño podemos aproximar el valor de la distancia de Wasserstein sin regularizar tanto como queramos.

Capítulo 3

Caso discreto

En este capítulo vamos a profundizar en los conceptos que hemos introducido en los dos capítulos anteriores en el caso particular de distribuciones discretas. Nos centraremos en distribuciones en un conjunto finito de puntos el cual denotaremos por $\mathcal{X} = \{x_1, \ldots, x_n\}$. La principal referencia que hemos seguido en este capítulo es [1].

Observación 10. Si consideramos un conjunto finito $\mathcal{X}$ y lo dotamos de la topología discreta el espacio topológico obtenido es polaco y compacto.

Notación 3.1. Sea $\mathcal{X}$ un conjunto finito sobre el cual está definida una probabilidad μ. Si fijamos un orden

$$(x_1, \ldots, x_n)$$

de los elementos de $\mathcal{X}$ podemos representar μ por un vector

$$\mathbf{m} = (m_1, \ldots, m_n)$$

donde $m_i = \mu(\{x_i\})$. De esta forma podemos escribir $\mu = \sum_{i=1}^{n} m_i \, \delta_{x_i}$.

Vamos a formular el problema de transporte óptimo en la versión de Kantorovich para el caso particular de dos distribuciones discretas. Para ello consideramos $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades sobre dos conjuntos finitos $\mathcal{X} = \{x_1, \ldots, x_m\}$ e $\mathcal{Y} = \{y_1, \ldots, y_n\}$. Consideraremos que los puntos de $\mathcal{X}$ e $\mathcal{Y}$ están ordenados y escribiremos

$$\alpha = \sum_{i=1}^{m} a_i \, \delta_{x_i} \quad \text{y} \quad \beta = \sum_{j=1}^{n} b_j \, \delta_{y_j}.$$

Emplearemos los vectores $\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^m$ y $\mathbf{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$ para referirnos a las probabilidades α y β respectivamente. En los casos de interés, la función de coste c que interviene en la definición del transporte óptimo solo tomará valores finitos. Por ello podemos observar que la función de coste c se puede representar como una matriz

$$\mathbf{C} = (c_{i,j}) \in \mathbb{R}^{m \times n}$$

donde $c_{i,j} \overset{\text{def.}}{=} c(x_i, y_j)$. En esta situación, decimos que $\mathbf{C}$ es la matriz de coste del problema de transporte óptimo.

Atendiendo a la notación que hemos introducido, podemos representar una distribución π en $\mathcal{X} \times \mathcal{Y}$ con marginales α y β de forma matricial: denotamos por $p_{i,j} = \pi(\{x_i, y_j\})$ para cada par (x_i, y_j) del espacio producto. La matriz positiva

$$\mathbf{P} = (p_{i,j}) \in \mathbb{R}^{m \times n}$$

recoge toda la información de la distribución π y satisface

$$\mathbf{P}\,\mathbb{1}_n = \mathbf{a} \qquad \text{es decir} \qquad \sum_{j=1}^{n} p_{i,j} = a_i\,,\ 1 \le i \le m \tag{3.1}$$

$$\mathbf{P}^T\,\mathbb{1}_m = \mathbf{b} \qquad \text{es decir} \qquad \sum_{i=1}^{m} p_{i,j} = b_j\,,\ 1 \le j \le n.$$

Emplearemos la notación que hemos introducido cuando trabajemos con distribuciones de probabilidad con soporte finito. En este contexto, introducimos el conjunto de matrices que son factibles en el problema de transporte óptimo discreto.

Definición 3.0.1. Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades sobre dos conjuntos finitos $\mathcal{X} = \{x_1, \dots, x_m\}$ e $\mathcal{Y} = \{y_1, \dots, y_n\}$, los cuales suponemos que están ordenados. El conjunto $\Pi(\alpha, \beta)$ se identifica con el conjunto de matrices

$$\mathcal{U}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \{\mathbf{P} \in \mathbb{R}^{m\times n}_{\ge 0} \,:\, \mathbf{P}\,\mathbb{1}_n = \mathbf{a},\ \mathbf{P}^T\,\mathbb{1}_m = \mathbf{b}\}.$$

Notación 3.2. Sean $\mathbf{P} = (p_{i,j})$ y $\mathbf{Q} = (q_{i,j})$ dos matrices reales de orden $m \times n$. Escribimos

$$\langle \mathbf{P}, \mathbf{Q} \rangle \overset{\text{def.}}{=} \sum_{i,j} p_{i,j} \cdot q_{i,j}.$$

Con las notaciones introducidas podemos reescribir el problema de Kantorovich en el caso discreto empleando lenguaje matricial. Esta formulación se corresponde con un problema de programación lineal de minimización ya que tanto la función objetivo como las restricciones son lineales.

Definición 3.0.2. Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades sobre dos espacios finitos y sea $\mathbf{C} \in \mathbb{R}^{m\times n}$ una matriz de coste. Denotamos por $\mathbf{a}$ y $\mathbf{b}$ a los vectores que representan a α y β respectivamente. El problema de transporte óptimo discreto se escribe

$$\mathcal{L}_c(\alpha, \beta) = \mathcal{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P}\in\mathcal{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle. \tag{3.2}$$

El problema de transporte óptimo siempre tiene solución en el caso discreto como se puede comprobar aplicando el Teorema 1.3.4 y la Observación 10, por ello podemos sustituir el inferior por el mínimo en la expresión (3.2). Este hecho también se puede ver como consecuencia de que en el caso discreto el problema de transporte óptimo es un problema de programación lineal factible: considerando la matriz $\mathbf{P} = (p_{i,j})$ definida por

$$p_{i,j} = a_i \cdot b_j \quad 1 \le i \le m,\ 1 \le j \le n \tag{3.3}$$

podemos ver que $\mathbf{P}$ es una solución factible del problema. Podemos observar que la matriz $\mathbf{P}$ que hemos definido es la que se corresponde con la distribución de probabilidad $\alpha \otimes \beta$.

Considerar el problema de transporte óptimo discreto como un problema de programación lineal nos permite emplear las técnicas de este campo para obtener una solución óptima. Es posible resolverlo con una variante del método Simplex denominada algoritmo húngaro, el cual está especializado en el problema de transporte óptimo. El inconveniente de esta herramienta es que la complejidad del algoritmo crece con el cubo del tamaño del problema lo cual supone un problema de escalabilidad.

Observación 11. El problema de transporte óptimo discreto se puede estudiar en contexto puramente vectorial, es decir, sin hacer mención a los espacios $\mathcal{X}$ e $\mathcal{Y}$ ni a las probabilidades α y β: fijada una matriz de coste $\mathbf{C} \in \mathbb{R}^{m\times n}$, podemos calcular el coste de transporte óptico entre dos vectores $\mathbf{a} \in \mathbb{R}^m$ y $\mathbf{b} \in \mathbb{R}^n$ como el valor del mínimo (3.2). La ventaja de esta formulación vectorial es que nos permite trabajar en con una notación más precisa y hacer énfasis en que se está trabajando en el caso discreto.

La observación previa nos permite trabajar con vectores y matrices en problemas de transporte óptico discreto. Debemos tener en cuenta que un vector $\mathbf{a} \in \mathbb{R}^m$ representa una probabilidad si y solo si la suma de sus componentes es 1. Atendiendo a esta consideración introducimos la siguiente definición.

Definición 3.0.3. Fijado un $m \geq 1$, se define el conjunto $\Sigma_n = \{\mathbf{a} = (a_1, \ldots, a_m) \in \mathbb{R}^n : \sum_{i=1}^m a_i = 1\}$. Este es el símplex n-dimensional.

Escribimos ahora la formulación del problema de transporte entrópico en el caso discreto haciendo uso de las notaciones que se han introducido.

Proposición 3.0.1 (Problema del transporte entrópico discreto corregido)**.** *Sean $\alpha \in \mathcal{P}(\mathcal{X})$ y $\beta \in \mathcal{P}(\mathcal{Y})$ dos probabilidades sobre dos espacios finitos las cuales se representan por los vectores $\mathbf{a} \in \mathbb{R}^m$ y $\mathbf{b} \in \mathbb{R}^n$. Sea $\mathbf{C} \in \mathbb{R}^{m \times n}$ una matriz de coste y $\varepsilon > 0$ el factor de regularización. El problema de transporte entrópico discreto corregido se escribe*

$$\min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle + \varepsilon \left[\sum_{i,j} h(p_{i,j}) - \sum_{i,j} h(a_i\, b_j) \right] \tag{3.4}$$

donde $h(z) = z \log(z)$ (se considera $h(0) = 0$).

Demostración. Para verificar que la expresión (3.4) se satisface, consideramos una distribución $\pi \in \Pi(\alpha, \beta)$ con matriz asociada $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$. Podemos comprobar que la entropía relativa de π respecto de $\alpha \otimes \beta$ se puede escribir

$$\mathrm{KL}\left(\pi | \alpha \otimes \beta\right) = \begin{cases} \infty & \text{si existen } i, j \text{ tales que } a_i\, b_j = 0 \text{ y } p_{i,j} > 0, \\ \displaystyle\sum_{i,j} \log\left(\frac{d\pi}{d\alpha \otimes \beta}(x_i, y_j) \right) p_{i,j} & \text{en otro caso.} \end{cases}$$

Esto se debe a que π no es absolutamente continua respecto de $\alpha \otimes \beta$ si y solo si existe un par (i, j) con $a_i\, b_j = 0$ y $p_{i,j} > 0$. Vamos a probar esta afirmación: supongamos que $a_{i_0} = 0$ y $p_{i_0, j_0} > 0$ y supongamos que π no es absolutamente continua respecto de $\alpha \otimes \beta$. Dado que $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ se tendría

$$0 = a_{i_0} = \sum_j p_{i,j} \geq p_{i_0, j_0} > 0$$

lo que es absurdo. De forma análoga se prueba que no puede darse $b_{i_0} = 0$ y $p_{i_0, j} > 0$.

Además, si π es absolutamente continua respecto de $\alpha \otimes \beta$ se tiene que la derivada de Radon-Nykodym de π respecto de $\alpha \otimes \beta$ es

$$\frac{d\pi}{d\alpha \otimes \beta}(x_i, y_j) = \begin{cases} 1 & \text{si } a_i\, b_j = 0 \\ \dfrac{p_{i,j}}{a_i\, b_j} & \text{en otro caso} \end{cases}$$

con lo que se deduce que $\mathrm{KL}\left(\pi | \alpha \otimes \beta\right) = \sum_{i,j} \log\left(\frac{p_{i,j}}{a_i\, b_j} \right) p_{i,j}$.

Desarrollando el sumatorio obtenemos

$$\begin{aligned} \sum_{i,j} \log\left(\frac{p_{i,j}}{a_i\, b_j} \right) p_{i,j} &= \sum_{i,j} \log\left(p_{i,j}\right) p_{i,j} - \sum_{i,j} \log\left(a_i\right) p_{i,j} - \sum_{i,j} \log\left(b_j\right) p_{i,j} \\ &= \sum_{i,j} \log\left(p_{i,j}\right) p_{i,j} - \sum_{i} \log\left(a_i\right) a_i - \sum_{j} \log\left(b_j\right) b_j \\ &= \sum_{i,j} h(p_{i,j}) - \sum_{i} h(a_i) - \sum_{j} h(b_j). \end{aligned}$$

Para finalizar la demostración, basta comprobar que $\sum_i h(a_i) + \sum_j h(b_j) = \sum_{i,j} h(a_i\, b_j)$:

$$\begin{aligned}\sum_{i,j} h(a_i\, b_j) &= \sum_{i,j} a_i b_j \log(a_i b_j) = \sum_{i,j} a_i b_j (\log(a_i) + \log(b_j)) \\ &= \sum_{i,j} a_i b_j \log(a_i) + \sum_{i,j} a_i b_j \log(b_j) \\ &= \sum_i a_i \log(a_i) \left(\sum_j b_j \right) + \sum_j b_j \log(b_j) \left(\sum_i a_i \right) \\ &= \sum_i a_i \log(a_i) + \sum_j b_j \log(b_j) = \sum_i h(a_i) + \sum_j h(b_j)\end{aligned}$$

□

Observación 12. Notamos que con las condiciones dadas en la Proposición 3.0.1 para el problema del transporte entrópico discreto corregido, se satisfacen las hipótesis del Teorema 2.0.5. Por lo tanto, sabemos que existe una única distribución $\pi^\varepsilon \in \Pi(\alpha, \beta)$ que es solución del problema. La probabilidad π^ε está asociada a la única matriz de $\mathcal{U}(\mathbf{a}, \mathbf{b})$ que minimiza (3.4), la cual denotaremos por $\mathbf{P}^\varepsilon$.

Podemos observar que el término $\sum_{i,j} h(a_i\, b_j)$ que interviene en la formulación del problema de transporte entrópico discreto es fijo y no depende de $\mathbf{P}$. Por lo tanto, una matriz $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ minimiza (3.4) si y solo si $\mathbf{P}$ minimiza

$$\langle \mathbf{C}, \mathbf{P} \rangle + \varepsilon \sum_{i,j} \log(p_{i,j}) p_{i,j}\,,$$

o equivalentemente si en $\mathbf{P}$ se alcanza el mínimo

$$\mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle + \varepsilon \sum_{i,j} p_{i,j} (\log(p_{i,j}) - 1). \tag{3.5}$$

Vamos a emplear esta última caracterización que hemos dado para obtener $\mathbf{P}^\varepsilon$. Diremos que $\mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ es el coste de transporte entrópico discreto entre $\mathbf{a}$ y $\mathbf{b}$. Podemos relacionarlo con el coste de transporte entrópico discreto corregido (3.4) por la siguiente igualdad:

$$\mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) + \varepsilon \left(1 - \sum_{i,j} h(a_i\, b_j) \right) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle + \varepsilon \left[\sum_{i,j} h(p_{i,j}) - \sum_{i,j} h(a_i\, b_j) \right] \tag{3.6}$$

Observación 13. La igualdad previa nos permite deducir que coste de transporte entrópico discreto converge al coste de transporte óptimo discreto cuando ε tiende a 0. La Proposición 2.0.6 implica que el coste de transporte entrópico discreto corregido es estrictamente convexo respecto de $(\mathbf{a}, \mathbf{b})$. Además, el término $-\varepsilon(1 - \sum_{i,j} h(a_i\, b_j))$ es convexo, por lo tanto la función $(\mathbf{a}, \mathbf{b}) \mapsto \mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ es estrictamente convexa por ser suma de una función convexa y una función estrictamente convexa.

Notación 3.3. Dada una matriz $\mathbf{P} \in \mathbb{R}^{m \times n}_{>0}$. Escribimos $\log(\mathbf{P})$ para referirnos a la matriz cuyas componentes son los logaritmos de las componentes correspondientes de $\mathbf{P}$. De forma análoga, escribiremos $e^{\mathbf{P}}$ para denotar a la matriz cuyas componentes son la exponencial de las componentes correspondientes de $\mathbf{P}$. Emplearemos la misma notación para vectores de $\mathbb{R}^k$. De esta forma el problema de transporte entrópico discreto se puede escribir de la siguiente manera:

$$\mathcal{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \langle \mathbf{P}, \log(\mathbf{P}) - \mathbb{1}_{m \times n} \rangle.$$

La siguiente proposición nos da una descomposición de de la matriz óptima del problema de transporte entrópico la cual vamos a explotar con el Algoritmo de Sinkhorn para calcular $\mathbf{P}^\varepsilon$.

Proposición 3.0.2. *Sean $\mathbf{a} \in \Sigma_m$ y $\mathbf{b} \in \Sigma_n$ dos vectores de probabilidades y sea $\mathbf{C} \in \mathbb{R}^{m\times n}$ una matriz de coste. Sea $\varepsilon > 0$ el factor de regularización. Existen dos vectores no negativos $\mathbf{u} \in \mathbb{R}^m_{\geq 0}$ y $\mathbf{v} \in \mathbb{R}^n_{\geq 0}$ tales que*

$$p^{\varepsilon}_{i,j} = u_i \cdot e^{-\frac{c_{i,j}}{\varepsilon}} \cdot v_j \tag{3.7}$$

donde $p^{\varepsilon}_{i,j}$ es la entrada (i,j) de la matriz óptima del problema del transporte entrópico discreto, $\mathbf{P}^{\varepsilon}$. Es decir, se tiene la descomposición $\mathbf{P}^{\varepsilon} = diag(\mathbf{u}) \cdot \mathcal{K}_{\varepsilon} \cdot diag(\mathbf{v})$, donde $\mathcal{K}_{\varepsilon}$ es la matriz de $\mathbb{R}^{m\times n}$ cuya componente (i,j)-ésima es $e^{-\frac{c_{i,j}}{\varepsilon}}$. Decimos que esta es la matriz de Gibbs.

Demostración. En primer lugar, podemos deducir que si $\mathbf{u}$ y $\mathbf{v}$ son dos vectores que satisfacen (3.7) entonces estos deben satisfacer:

$$\begin{cases} u_i = 0 & \text{si } a_i = 0 \\ v_j = 0 & \text{si } b_j = 0. \end{cases}$$

Esto se debe a que si $a_i = 0$ para un cierto $1 \leq i \leq m$, entonces $p^{\varepsilon}_{i,j} = 0$ para cada $1 \leq j \leq n$. Si se tuviese $u_i > 0$, entonces para cada $1 \leq j \leq n$ deduciríamos que $v_j e^{-\frac{c_{i,j}}{\varepsilon}} = 0$ y por lo tanto $v_j = 0$. Pero en esta situación se tendría que la matriz $\mathbf{P}^{\varepsilon}$ sería nula lo cual no puede suceder. De una forma análoga se prueba que v_j si $b_j = 0$.

Trabajando en los soportes de $\mathbf{a}$ y $\mathbf{b}$ (es decir, restringiéndonos a los soportes de las probabilidades α, β asociadas) podemos suponer que $a_i > 0$ para cada $1 \leq i \leq m$ y $b_j > 0$ para cada $1 \leq j \leq n$. En esta situación, la matriz $\mathbf{P}^{\varepsilon}$ es estrictamente positiva y en consecuencia los vectores $\mathbf{u}$ y $\mathbf{v}$ son estrictamente positivos. Vamos a aplicar el método de los multiplicadores de Lagrange para obtener la única matriz en la que se alcanza el mínimo (3.5). Para ello notamos que tenemos $m + n$ restricciones: las m primeras se deben a la condición $\mathbf{P}\mathbb{1}_n = \mathbf{a}$, y las n restantes se obtienen de la condición $\mathbf{P}^T\mathbb{1}_m = \mathbf{b}$. Para cada $1 \leq i \leq m$ consideramos el multiplicador f_i asociado a la restricción $\sum_j p_{i,j} = a_i$. De forma análoga, para cada $1 \leq j \leq n$ consideramos el multiplicador g_j asociado a la restricción $\sum_i p_{i,j} = b_i$.

Definimos la función Lagrangiana del problema de transporte entrópico discreto (3.5)

$$\mathfrak{L}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{C}, \mathbf{P} \rangle + \varepsilon \sum_{i,j} p_{i,j}(\log(p_{i,j}) - 1) - \sum_i f_i \cdot \left(\sum_j p_{i,j} - a_i \right) - \sum_j g_j \cdot \left(\sum_i p_{i,j} - b_j \right).$$

La única matriz de $\mathcal{U}(\mathbf{a}, \mathbf{b})$ en la que se alcanza $\mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$, que hemos denotado por $\mathbf{P}^{\varepsilon}$, debe verificar

$$\frac{\partial \mathfrak{L}}{\partial p_{i,j}}(\mathbf{P}^{\varepsilon}, \mathbf{f}, \mathbf{g}) = 0$$

para cada $1 \leq i \leq m$ y $1 \leq j \leq n$ para unos ciertos vectores $\mathbf{f} = (f_1, \ldots, f_m)$ y $\mathbf{g} = (g_1, \ldots, g_n)$. La derivada parcial de $\mathfrak{L}$ respecto de $p_{i,j}$ es

$$\frac{\partial \mathfrak{L}}{\partial p_{i,j}} = c_{i,j} + \varepsilon \, \log(p_{i,j}) - f_i - g_j. \tag{3.8}$$

Para cada $1 \leq i \leq m$ y $1 \leq j \leq n$ imponemos que la expresión (3.8) se anule, lo que nos permite despejar $p_{i,j}$:

$$p_{i,j} = e^{\frac{(f_i - c_{i,j} + g_j)}{\varepsilon}} = e^{\frac{f_i}{\varepsilon}} \cdot e^{-\frac{c_{i,j}}{\varepsilon}} \cdot e^{\frac{g_j}{\varepsilon}}. \tag{3.9}$$

Denotando por $u_i = e^{\frac{f_i}{\varepsilon}}$ y $v_j = e^{\frac{g_j}{\varepsilon}}$ se concluye el resultado. □

Observación 14. Los vectores $\mathbf{u}$ y $\mathbf{v}$ que intervienen en la Proposición 3.0.2 no son únicos ya para cada escalar positivo λ los vectores $\lambda\mathbf{u}$ y $\frac{1}{\lambda}\mathbf{v}$ también satisfacen la condición (3.7). Sin embargo, los pares de vectores de la forma

$$\left(\lambda\mathbf{u}, \tfrac{1}{\lambda}\mathbf{v}\right)$$

para $\lambda > 0$ son los únicos que satisfacen esta condición. En caso contrario, existirían un par de vectores $(\mathbf{u}', \mathbf{v}')$ satisfaciendo (3.7) e índices $1 \leq i_1, i_2 \leq m$ distintos tales que

$$\frac{u'_{i_1}}{u_{i_1}} \neq \frac{u'_{i_2}}{u_{i_2}}.$$

Dado que cada par de vectores satisface (3.7), podemos deducir que $u_{i_1} v_j = u'_{i_1} v'_j$ y $u_{i_2} v_j = u'_{i_2} v'_j$, de donde despejamos

$$v'_j = v_j \frac{u_{i_1}}{u'_{i_1}} \quad y \quad v'_j = v_j \frac{u_{i_2}}{u'_{i_2}}$$

para un índice j que satisfaga $v_j > 0$. Entonces deducimos que

$$v_j \frac{u_{i_1}}{u'_{i_1}} - v_j \frac{u_{i_2}}{u'_{i_2}} = v_j \left(\frac{u'_{i_1}}{u_{i_1}} - \frac{u'_{i_2}}{u_{i_2}} \right) = 0$$

llegando a una contradicción.

Observación 15. Como ya se ha comprobado en la Proposición 3.0.2, un elemento $p_{i,j}$ de la matriz $\mathbf{P}^\varepsilon$ es 0 si y solo si se tiene que la fila i-ésima o la columna j-ésima de $\mathbf{P}^\varepsilon$ son nulas. Además, el primer caso ocurre solo cuando el coeficiente a_i es 0. De igual forma, la columna j-ésima de la matriz $\mathbf{P}^\varepsilon$ es nula solo si el coeficiente b_j es 0. Es decir, el soporte de la solución del problema de transporte entrópico discreto es igual al producto de los soportes de α y β:

$$\text{supp}(\pi_\varepsilon) = \text{supp}(\alpha) \times \text{supp}(\beta).$$

Notación 3.4. La Observación15 nos permite trabajar únicamente con los soportes de α y β a los que denotaremos por $\mathcal{X}$ e $\mathcal{Y}$ para aligerar la notación. De esta forma, podemos asumir que los vectores de probabilidades $\mathbf{a}$ y $\mathbf{b}$ son estrictamente positivos.

3.1. Algoritmo de Sinkhorn

En esta sección se va a introducir un método numérico que permite obtener la matriz de transporte entrópico en el caso discreto. El Algoritmo de Sinkhorn goza de buenas propiedades porque es fácil de implementar y se puede paralelizar lo que supone una ventaja a nivel computacional frente a los algoritmos que se emplean para obtener una solución del problema no regularizado. Previamente introduciremos la distancia proyectiva de Hilbert y daremos algunas propiedades de esta que nos permitan probar la convergencia del algoritmo.

Definición 3.1.1. Sean $\mathbf{u}, \mathbf{v}$ dos vectores estrictamente positivos de $\mathbb{R}^n$. La distancia proyectiva de Hilbert entre los vectores está dada por

$$d_{\mathcal{H}}(\mathbf{u}, \mathbf{v}) \overset{\text{def.}}{=} \log \left(\max_{i,j} \frac{u_i \, v_j}{u_j \, v_i} \right). \tag{3.10}$$

Proposición 3.1.1. *Se define el cono proyectivo $\mathbb{R}^n_{>0}/\sim$ donde $\mathbf{u} \sim \mathbf{v}$ si y solo si existe un escalar $\lambda > 0$ tal que $\mathbf{u} = \lambda \mathbf{v}$. Es decir, en $\mathbb{R}^n_{>0}/\sim$ se identifican vectores que son múltiplos positivos. La distancia proyectiva de Hilbert es una distancia en $\mathbb{R}^n_{>0}/\sim$.*

Demostración. En primer lugar podemos comprobar que la distancia proyectiva de Hilbert está bien definida en el cono proyectivo ya que basta notar que se verifica

$$\frac{u_i \, v_j}{u_j \, v_i} = \frac{(\lambda u_i) \cdot (\mu v_j)}{(\lambda u_j) \cdot (\mu v_i)}$$

para cada $\lambda, \mu > 0$. Además, dados dos vectores $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n_{>0}$ podemos observar que

$$\max_{i,j} \frac{u_i \, v_j}{u_j v_i}$$

es mayor o igual que 1. En caso contrario, tomando los índices i_0, j_0 en los que se alcanza el máximo se tendría

$$\max_{i,j} \frac{u_i\, v_j}{u_j\, v_i} = \frac{u_{i_0}\, v_{j_0}}{u_{j_0}\, v_{i_0}} < 1 < \frac{u_{j_0}\, v_{i_0}}{u_{i_0}\, v_{j_0}}$$

lo que supondría una contradicción. De esta forma, se comprueba que la distancia de Hilbert entre dos elementos del cono proyectivo esta bien definida y es siempre positiva. Por la inyectividad del logaritmo se tiene que $d_{\mathcal{H}}(\mathbf{u}, \mathbf{v}) = 0$ si

$$\max_{i,j} \tfrac{u_i\, v_j}{u_j\, v_i} = \left(\max_i \tfrac{u_i}{v_i}\right)\left(\max_j \tfrac{v_j}{u_j}\right) = 1,$$

de donde se deduce que

$$\max_i \tfrac{u_i}{v_i} = \left(\max_j \tfrac{v_j}{u_j}\right)^{-1} = \min_j \tfrac{u_j}{v_j},$$

lo cual solo ocurre si $\mathbf{u} = \left(\max_i \frac{u_i}{v_i}\right) \cdot \mathbf{v}$, es decir, si $\mathbf{u}$ y $\mathbf{v}$ pertenecen a la misma clase de equivalencia de $\mathbb{R}^n_{>0}/\sim$. Se puede ver claramente que $d_{\mathcal{H}}$ es simétrica, por lo que basta probar la desigualdad triangular para probar que efectivamente es una distancia. Consideramos $\mathbf{u}, \mathbf{v}, w \in \mathbb{R}^n$. Se tiene que

$$\max_{i,j} \frac{u_i\, v_j}{u_j\, v_i} = \max_{i,j} \left(\frac{u_i\, w_j}{u_j\, w_i} \cdot \frac{w_i\, v_j}{w_j\, v_i}\right) \leq \max_{i,j} \left(\frac{u_i\, w_j}{u_j\, w_i}\right) \cdot \max_{i,j} \left(\frac{w_i\, v_j}{w_j\, v_i}\right).$$

Tomando logaritmos en esta expresión deducimos la desigualdad triangular. □

La distancia proyectiva de Hilbert nos proporciona el marco adecuado para estudiar la convergencia del Algoritmo de Sinkhorn. La prueba está basada en el siguiente Teorema de Birkhoff.

Teorema 3.1.1 (Birkhoff)**.** *Sea $\mathbf{K} \in \mathbb{R}^{m\times n}$ una matriz estrictamente positiva. Entonces la premultiplicación de un vector de $\mathbb{R}^m_{>0}$ por $\mathbf{K}$ es una aplicación contractiva. Es decir,*

$$d_{\mathcal{H}}(\mathbf{K}\mathbf{u}, \mathbf{K}\mathbf{v}) \leq \lambda(\mathbf{K})\, d_{\mathcal{H}}(\mathbf{u}, \mathbf{v}) \tag{3.11}$$

donde se definen

$$\begin{cases} \lambda(\mathbf{K}) = \dfrac{\sqrt{\eta(\mathbf{K})} - 1}{\sqrt{\eta(\mathbf{K})} + 1} < 1 \\ \eta(\mathbf{K}) = \max\limits_{i,j,k,l} \dfrac{\mathbf{K}_{i,k}\, \mathbf{K}_{j,l}}{\mathbf{K}_{j,k}\, \mathbf{K}_{i,l}}. \end{cases}$$

Además, esta cota es óptima en el sentido de que para cada $0 < \tilde{\lambda} < \lambda(\mathbf{K})$, existe un par de vectores positivos $\tilde{\mathbf{u}}, \tilde{\mathbf{v}}$ para los cuales

$$d_{\mathcal{H}}(\mathbf{K}\tilde{\mathbf{u}}, \mathbf{K}\tilde{\mathbf{v}}) > \tilde{\lambda}\, d_{\mathcal{H}}(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}).$$

Demostración. La demostración de este resultado es una consecuencia de [10, Thm.1]. □

Lema 3.1.2. Sea $\mathbf{K} \in \mathbb{R}^{m\times n}$ una matriz estrictamente positiva. Sea $\lambda(\mathbf{K})$ definido como en el Teorema 3.1.1. Entonces $\lambda(\mathbf{K}) = \lambda(\mathbf{K}^T)$.

Demostración. La igualdad $\lambda(\mathbf{K}) = \lambda(\mathbf{K}^T)$ es consecuencia de

$$\eta(\mathbf{K}^T) = \max_{i,j,k,l} \frac{\mathbf{K}^T_{i,k}\, \mathbf{K}^T_{j,l}}{\mathbf{K}^T_{j,k}\, \mathbf{K}^T_{i,l}} = \max_{i,j,k,l} \frac{\mathbf{K}_{k,i}\, \mathbf{K}_{l,j}}{\mathbf{K}_{k,j}\, \mathbf{K}_{l,i}} = \eta(\mathbf{K}).$$

□

Notación 3.5. Dados dos vectores $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n_{>0}$, escribimos $\frac{\mathbf{u}}{\mathbf{v}}$ para denotar al vector de $R^n_{>0}$ que tiene $\frac{u_i}{v_i}$ por componente i-ésima. También utilizaremos la notación $\mathbf{u} \odot \mathbf{v}$ para designar el vector dado por la multiplicación componente a componente de $\mathbf{u}$ y $\mathbf{v}$.

Lema 3.1.3. Sean $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n_{>0}$. Se satisface $d_{\mathcal{H}}(\mathbf{u}, \mathbf{v}) = d_{\mathcal{H}}\left(\frac{\mathbf{u}}{\mathbf{v}}, \mathbb{1}\right)$.

Demostración. Basta comprobar que $d_{\mathcal{H}}\left(\frac{\mathbf{u}}{\mathbf{v}}, \mathbb{1}\right)$ se puede escribir de la siguiente manera:

$$d_{\mathcal{H}}\left(\frac{\mathbf{u}}{\mathbf{v}}, \mathbb{1}\right) = \log\left(\max_{i,j} \frac{u_i/v_i \cdot 1}{1 \cdot u_j/v_j}\right) = \log\left(\max_{i,j} \frac{u_i\, v_j}{v_i\, u_j}\right).$$

□

El Teorema de Birkhoff (Teorema 3.1.1) nos permite dar una caracterización de la solución del problema de transporte entrópico discreto a partir de la descomposición (3.7).

Proposición 3.1.2 (Caracterización de $\mathbf{P}^\varepsilon$)**.** *Consideramos un problema de transporte entrópico discreto* (3.5) *con vectores* $\mathbf{a}$ *y* $\mathbf{b}$ *estrictamente positivos. Sea* $\mathbf{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ *una matriz para la cual existen vectores* $\mathbf{u} \in \mathbb{R}^m_{>0}$ *y* $\mathbf{v} \in \mathbb{R}^n_{>0}$ *tales que se da la descomposición*

$$\mathbf{P} = diag(\mathbf{u}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}).$$

Entonces $\mathbf{P}$ *es la solución del problema de transporte entrópico discreto, es decir,* $\mathbf{P} = \mathbf{P}^\varepsilon$.

Demostración. Sea $\mathbf{P}^\varepsilon$ la solución del problema de transporte entrópico discreto considerado. Consideramos los vectores $\mathbf{u}^\varepsilon \in \mathbb{R}^m_{>0}$ y $\mathbf{v}^\varepsilon \in \mathbb{R}^n_{>0}$ dados por la Proposición 3.0.2, los cuales satisfacen

$$\mathbf{P}^\varepsilon = diag(\mathbf{u}^\varepsilon) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}^\varepsilon).$$

Dado que la matriz $\mathbf{P}$ pertenece a $\mathcal{U}(\mathbf{a}, \mathbf{b})$ se verifican las igualdades

$$\begin{aligned} \mathbf{a} &= \mathbf{P}\mathbb{1} = diag(\mathbf{u}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v})\mathbb{1} = \mathbf{u} \odot \mathcal{K}_\varepsilon \mathbf{v} \\ \mathbf{b} &= \mathbf{P}^T\mathbb{1} = diag(\mathbf{v}) \cdot \mathcal{K}_\varepsilon^T \cdot diag(\mathbf{u})\mathbb{1} = \mathbf{u} \odot \mathcal{K}_\varepsilon^T \mathbf{u}, \end{aligned} \tag{3.12}$$

de donde se deduce $\mathbf{u} = \frac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}}$ y $\mathbf{v} = \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}}$. De forma similar, dado que $\mathbf{P}^\varepsilon \in \mathcal{U}(\mathbf{a}, \mathbf{b})$ se tiene

$$\begin{aligned} \mathbf{a} &= \mathbf{P}^\varepsilon\mathbb{1} = diag(\mathbf{u}^\varepsilon) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}^\varepsilon)\mathbb{1} = \mathbf{u}^\varepsilon \odot \mathcal{K}_\varepsilon \mathbf{v}^\varepsilon \\ \mathbf{b} &= (\mathbf{P}^\varepsilon)^T\mathbb{1} = diag(\mathbf{v}^\varepsilon) \cdot \mathcal{K}_\varepsilon^T \cdot diag(\mathbf{u}^\varepsilon)\mathbb{1} = \mathbf{u}^\varepsilon \odot \mathcal{K}_\varepsilon^T \mathbf{v}^\varepsilon, \end{aligned} \tag{3.13}$$

de donde se deduce $\mathbf{u}^\varepsilon = \frac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}^\varepsilon}$ y $\mathbf{v}^\varepsilon = \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}^\varepsilon}$. Dividiendo las expresiones de $\mathbf{u}^\varepsilon$ y $\mathbf{v}^\varepsilon$ entre las expresiones que hemos obtenido para $\mathbf{u}$ y $\mathbf{v}$ obtenemos

$$\frac{\mathbf{u}^\varepsilon}{\mathbf{u}} = \frac{\mathcal{K}_\varepsilon \mathbf{v}}{\mathcal{K}_\varepsilon \mathbf{v}^\varepsilon} \quad \text{y} \quad \frac{\mathbf{v}^\varepsilon}{\mathbf{v}} = \frac{\mathcal{K}_\varepsilon^T \mathbf{u}}{\mathcal{K}_\varepsilon^T \mathbf{u}^\varepsilon}.$$

Entonces, aplicando los Lemas 3.1.3, 3.1.2 y el Teorema 3.1.1 se tiene

$$d_{\mathcal{H}}(\mathbf{u}^\varepsilon, \mathbf{u}) = d_{\mathcal{H}}\left(\frac{\mathbf{u}^\varepsilon}{\mathbf{u}}, \mathbb{1}\right) = d_{\mathcal{H}}\left(\frac{\mathcal{K}_\varepsilon \mathbf{v}}{\mathcal{K}_\varepsilon \mathbf{v}^\varepsilon}, \mathbb{1}\right) = d_{\mathcal{H}}\left(\mathcal{K}_\varepsilon \mathbf{v}, \mathcal{K}_\varepsilon \mathbf{v}^\varepsilon\right) \leq \lambda(\mathcal{K}_\varepsilon) d_{\mathcal{H}}(\mathbf{v}^\varepsilon, \mathbf{v})$$

$$d_{\mathcal{H}}(\mathbf{v}^\varepsilon, \mathbf{v}) = d_{\mathcal{H}}\left(\frac{\mathbf{v}^\varepsilon}{\mathbf{v}}, \mathbb{1}\right) = d_{\mathcal{H}}\left(\frac{\mathcal{K}_\varepsilon^T \mathbf{u}}{\mathcal{K}_\varepsilon^T \mathbf{u}^\varepsilon}, \mathbb{1}\right) = d_{\mathcal{H}}\left(\mathcal{K}_\varepsilon^T \mathbf{u}, \mathcal{K}_\varepsilon^T \mathbf{u}^\varepsilon\right) \leq \lambda(\mathcal{K}_\varepsilon^T) d_{\mathcal{H}}(\mathbf{u}^\varepsilon, \mathbf{u}) = \lambda(\mathcal{K}_\varepsilon) d_{\mathcal{H}}(\mathbf{u}^\varepsilon, \mathbf{u}),$$

donde $\lambda(\mathcal{K}_\varepsilon)$ está definida como en el Teorema 3.1.1. A partir de las dos expresiones anteriores se deduce

$$d_{\mathcal{H}}(\mathbf{u}^\varepsilon, \mathbf{u}) \leq \lambda(\mathcal{K}_\varepsilon)^2 d_{\mathcal{H}}(\mathbf{u}^\varepsilon, \mathbf{u}) \quad \text{y} \quad d_{\mathcal{H}}(\mathbf{v}^\varepsilon, \mathbf{v}) \leq \lambda(\mathcal{K}_\varepsilon)^2 d_{\mathcal{H}}(\mathbf{v}^\varepsilon, \mathbf{v}).$$

En consecuencia, por ser $\lambda(\mathcal{K}_\varepsilon)^2 < 1$ se deduce que $d_{\mathcal{H}}(\mathbf{u}^\varepsilon, \mathbf{u}) = 0$ y $d_{\mathcal{H}}(\mathbf{v}^\varepsilon, \mathbf{v}) = 0$. Esto indica que $\mathbf{u} = \lambda \mathbf{u}^\varepsilon$ y $\mathbf{v} = \mu \mathbf{v}^\varepsilon$ para dos escalares estrictamente positivos λ, μ. Además, por las igualdades (3.12) y (3.13) se deduce que $\lambda \cdot \mu = 1$; es decir, $\mathbf{P} = \mathbf{P}^\varepsilon$. □

Observación 16. La caracterización de la solución del problema de transporte entrópico discreto dada por la Proposición 3.1.2 nos permite transformar el problema (3.5) al siguiente problema: obtener dos vectores $\mathbf{u} \in \mathbb{R}^m_{>0}$ y $\mathbf{v} \in \mathbb{R}^n_{>0}$ tales que $diag(\mathbf{u}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}) \in \mathcal{U}(\mathbf{a}, \mathbf{b})$; es decir, que verifiquen $\mathbf{a} = \mathbf{u} \odot \mathcal{K}_\varepsilon \mathbf{v}$ y $\mathbf{b} = \mathbf{u} \odot \mathcal{K}_\varepsilon^T \mathbf{v}$, o equivalentemente

$$\mathbf{u} = \frac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}} \quad \text{y} \quad \mathbf{v} = \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}}. \tag{3.14}$$

Las Proposiciones 3.0.2 y 3.1.2 son un caso particular de un resultado general de la teoría de transporte entrópico. El Teorema 4.2 de [7] recoge este resultado el cual no hace uso de la distancia proyectiva de Hilbert para su demostración.

El Algoritmo de Sinkhorn, el cual se muestra en el Algoritmo 1, es un algoritmo de punto fijo que obtiene aproximaciones sucesivas de un par de vectores $\mathbf{u}^\varepsilon, \mathbf{v}^\varepsilon$ que satisfacen (3.7). Está basado en la reformulación del problema de transporte entrópico discreto que se ha dado en la Observación 16. En cada iteración del algoritmo se calculan los vectores $\mathbf{u}^{(l+1)} \in \mathbb{R}^m_{>0}$ y $\mathbf{v}^{(l+1)} \in \mathbb{R}^n_{>0}$ a partir de los iterantes previos y se obtiene una aproximación a la matriz $\mathbf{P}^\varepsilon$ la cual viene dada por

$$\mathbf{P}^{(l+1)} = diag(\mathbf{u}^{(l+1)}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}^{(l+1)}).$$

El cálculo de los iterantes $\mathbf{u}^{(l+1)}$ y $\mathbf{v}^{(l+1)}$ se lleva a cabo de la siguiente manera:

1. Se fija $\mathbf{v} = \mathbf{v}^{(l)}$ y se calcula el vector $\mathbf{u}^{(l+1)}$ de forma que este satisfaga la primera igualdad de (3.14).
2. Se fija $\mathbf{u}$ al vector $\mathbf{u}^{(l+1)}$ que se ha calculado en el paso previo y se obtiene $\mathbf{v}^{(l+1)}$ de forma que este verifique la segunda igualdad de (3.14).

Algoritmo 1: Algoritmo de Sinkhorn

Input: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \varepsilon, N$
Output: $\mathbf{u}, \mathbf{v}, \mathbf{P}$
$\mathbf{v}^{(0)} \leftarrow \mathbf{1}$;
while $l \leq N - 1$ **do**
  $\mathbf{u}^{(l+1)} \leftarrow \dfrac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}^{(l)}}$;
  $\mathbf{u}^{(l+1)} \leftarrow \dfrac{1}{\mathbf{u}_1^{(l+1)}} \mathbf{u}^{(l+1)}$;
  $\mathbf{v}^{(l+1)} \leftarrow \dfrac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}^{(l+1)}}$;
end
$\mathbf{u} \leftarrow \mathbf{u}^{(N)}$;
$\mathbf{v} \leftarrow \mathbf{v}^{(N)}$;
$\mathbf{P} \leftarrow diag(\mathbf{u}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v})$

Lema 3.1.4. Sean $\mathbf{u}^{(l)}, \mathbf{v}^{(l)}$ los vectores obtenidos por el Algoritmo de Sinkhorn en la iteración l-ésima y sea $\mathbf{P}^{(l)} = diag(\mathbf{u}^{(l)}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}^{(l)})$. Entonces

$$\mathbf{u}^{(l)} \odot \mathcal{K}_\varepsilon \mathbf{v}^{(l)} = \mathbf{P}^{(l)} \mathbb{1} \quad \text{y} \quad \mathcal{K}_\varepsilon^T \mathbf{u}^{(l)} \odot \mathbf{v}^{(l)} = \left(\mathbf{P}^{(l)}\right)^T \mathbb{1}.$$

Además, para cada $l \geq 1$ se verifica que $\left(\mathbf{P}^{(l)}\right)^T \mathbb{1} = \mathbf{b}$.

Demostración. Basta notar que

$$\mathbf{u}^{(l)} \odot \mathcal{K}_\varepsilon \mathbf{v}^{(l)} = \mathbf{u}^{(l)} \odot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}^{(l)})\mathbb{1} = diag(\mathbf{u}^{(l)}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v}^{(l)})\mathbb{1} = \mathbf{P}^{(l)}\mathbb{1},$$

$$\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)} \odot \mathbf{v}^{(l)} = \mathcal{K}_\varepsilon^T \cdot diag(\mathbf{u}^{(l)})\mathbb{1} \odot \mathbf{v}^{(l)} = diag(\mathbf{v}^{(l)}) \cdot \mathcal{K}_\varepsilon^T \cdot diag(\mathbf{u}^{(l)})\mathbb{1} = \left(\mathbf{P}^{(l)}\right)^T \mathbb{1}.$$

Además, por la definición de $\mathbf{v}^{(l)}$, podemos deducir que $\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)} \odot \mathbf{v}^{(l)} = \mathcal{K}_\varepsilon^T \mathbf{u}^{(l)} \odot \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)}} = \mathbf{b}$. □

Lema 3.1.5. Sean $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n_{>0}$. Se verifica

$$d_{\mathcal{H}}(\mathbf{u}, \mathbf{v}) \leq 2\| \log(\mathbf{u}) - \log(\mathbf{v})\|_\infty.$$

Además si $u_i = v_i$ para algún índice $1 \leq i \leq n$, también se tiene que

$$\| \log(\mathbf{u}) - \log(\mathbf{v})\|_\infty \leq d_{\mathcal{H}}(\mathbf{u}, \mathbf{v}).$$

Demostración.

$$\begin{aligned} d(\mathbf{u}, \mathbf{v}) &= \max_{i,j} \left(\log\left(u_i\, v_j\right) - \log\left(v_i\, u_j\right)\right) = \max_i \left(\log\left(u_i\right) - \log\left(v_i\right)\right) - \min_j \left(\log\left(u_j\right) - \log\left(v_j\right)\right) \\ &\leq \| \log(\mathbf{u}) - \log(\mathbf{v})\|_\infty + \| \log(\mathbf{u}) - \log(\mathbf{v})\|_\infty \end{aligned}$$

Además, si $u_{i_0} = v_{i_0}$ para algún índice $1 \leq i_0 \leq n$, se tiene que $\log(u_{i_0}) - \log(v_{i_0}) = 0$ y por lo tanto

$$d(\mathbf{u}, \mathbf{v}) = \max_i \left(\log\left(u_i\right) - \log\left(v_i\right)\right) - \min_j \left(\log\left(u_j\right) - \log\left(v_j\right)\right) \geq \max_i \left(\log\left(u_i\right) - \log\left(v_i\right)\right).$$

□

Observación 17. En particular, dados dos elementos $\overline{\mathbf{u}}, \overline{\mathbf{v}}$ del cono proyectivo $\mathbb{R}^n_{>0}/\sim$, podemos encontrar dos representantes suyos $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n_{>0}$ tales que

$$\| \log(\mathbf{u}) - \log(\mathbf{v})\|_\infty \leq d_{\mathcal{H}}(\overline{\mathbf{u}}, \overline{\mathbf{v}}) \leq 2\| \log(\mathbf{u}) - \log(\mathbf{v})\|_\infty.$$

Los resultados previos sobre la distancia proyectiva de Hilbert permiten demostrar el siguiente teorema, en el cual se prueba que el Algoritmo de Sinkhorn converge a la solución del problema de transporte entrópico discreto.

Teorema 3.1.6 (Convergencia del Algoritmo de Sinkhorn)**.** *Consideramos el problema del transporte entrópico discreto y suponemos que supp*$(\alpha) = \mathcal{X}$ *y supp*$(\beta) = \mathcal{Y}$*. Si* $\mathbf{u}, \mathbf{v}$ *son un par de vectores que satisfacen* (3.7) *y* $u_1 = 1$*, entonces* $(\mathbf{u}^{(l)}, \mathbf{v}^{(l)}) \longrightarrow (\mathbf{u}, \mathbf{v})$ *en la norma infinito y*

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) &= O\left(\lambda(\mathcal{K}_\varepsilon)^{2l}\right) \\ d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) &= O\left(\lambda(\mathcal{K}_\varepsilon)^{2l}\right), \end{aligned}$$

donde $\lambda(\mathcal{K}_\varepsilon)$ *está definida como en el Teorema 3.1.1. Además, se satisfacen las siguientes acotaciones*

$$d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) \leq \frac{d_{\mathcal{H}}\left(\mathbf{P}^{(l)}\mathbb{1}, \mathbf{a}\right)}{1 - \lambda(\mathcal{K}_\varepsilon)^2} \quad y \quad d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) \leq \frac{d_{\mathcal{H}}\left(\left(\mathbf{P}^{(l)}\right)^T \mathbb{1}, \mathbf{b}\right)}{1 - \lambda(\mathcal{K}_\varepsilon)^2}. \tag{3.15}$$

Demostración. Consideramos un par de vectores $\mathbf{u}, \mathbf{v}$ que satisfaga (3.7). Aplicando el Lema 3.1.3 repetidas veces, la Proposición 3.1.2 y el Teorema 3.1.1 podemos obtener las siguientes cadenas de desigualdades:

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(l+1)}, \mathbf{u}) &= d_{\mathcal{H}}\left(\frac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}^{(l)}}, \frac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}}\right) = d_{\mathcal{H}}\left(\frac{\mathbb{1}}{\mathcal{K}_\varepsilon \mathbf{v}^{(l)}}, \frac{\mathbb{1}}{\mathcal{K}_\varepsilon \mathbf{v}}\right) \\ &= d_{\mathcal{H}}\left(\mathcal{K}_\varepsilon \mathbf{v}^{(l)}, \mathcal{K}_\varepsilon \mathbf{v}\right) \leq \lambda(\mathcal{K}_\varepsilon)\, d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}), \end{aligned}$$

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) &= d_{\mathcal{H}}\left(\frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)}}, \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}}\right) = d_{\mathcal{H}}\left(\frac{\mathbb{1}}{\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)}}, \frac{\mathbb{1}}{\mathcal{K}_\varepsilon^T \mathbf{u}}\right) \\ &= d_{\mathcal{H}}\left(\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)}, \mathcal{K}_\varepsilon^T \mathbf{u}\right) \leq \lambda(\mathcal{K}_\varepsilon^T)\, d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) = \lambda(\mathcal{K}_\varepsilon)\, d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}). \end{aligned}$$

Aplicando estas desigualdades de forma recurrente y notando que $\lambda(\mathcal{K}_\varepsilon) < 1$ se deduce

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) &\leq \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{v}^{(l-1)}, \mathbf{v}) \leq \lambda(\mathcal{K}_\varepsilon)^{2l}\, d_{\mathcal{H}}(\mathbf{v}^{(0)}, \mathbf{v}) = O\left(\lambda(\mathcal{K}_\varepsilon)^{2l}\right), \\ d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) &\leq \lambda(\mathcal{K}_\varepsilon)\, d_{\mathcal{H}}(\mathbf{v}^{(l-1)}, \mathbf{v}) \leq \lambda(\mathcal{K}_\varepsilon)^{2l-1}\, d_{\mathcal{H}}(\mathbf{v}^{(0)}, \mathbf{v}) = O\left(\lambda(\mathcal{K}_\varepsilon)^{2l}\right). \end{aligned}$$

En particular se tiene que la clase proyectiva de $\mathbf{u}^{(l)}$ converge a la clase de $\mathbf{u}$ en $\mathbb{R}^n_{>0}/\sim$ y de igual forma deducimos que la clase proyectiva de $\mathbf{v}^{(l)}$ converge a la clase de $\mathbf{v}$ en $\mathbb{R}^n_{>0}/\sim$.

Podemos observar, que para cada $l \geq 1$ se tiene $\mathbf{u}_1^{(l)} = 1 = u_1$ por definición. Por esta razón, aplicando el Lema 3.1.5 sabemos que $\|\mathbf{u}^{(l)} - \mathbf{u}\|_\infty \leq d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u})$ de donde deducimos que $\lim\limits_{l\to\infty} \mathbf{u}^{(l)} = \mathbf{u}$ en la norma infinito. Es decir, cada componente de $\mathbf{u}^{(l)}$ converge a la correspondiente componente del vector $\mathbf{u}$. Ahora tomando el límite en la definición de $\mathbf{v}^{(l)}$ obtenemos

$$\mathbf{v}^{(l)} = \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}^{(l)}} \longrightarrow \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}} = \mathbf{v}.$$

En consecuencia, se tiene que $\lim\limits_{l\to\infty} (\mathbf{u}^{(l)}, \mathbf{v}^{(l)}) = (\mathbf{u}, \mathbf{v})$ en la norma infinito (o en cualquier otra norma equivalente).

Para probar las expresiones (3.15), vamos a emplear la desigualdad triangular y el Lema 3.1.4. De esta forma obtenemos

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) &\leq d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}^{(l+1)}) + d_{\mathcal{H}}(\mathbf{u}^{(l+1)}, \mathbf{u}) \leq d_{\mathcal{H}}\left(\mathbf{u}^{(l)}, \frac{\mathbf{a}}{\mathcal{K}_\varepsilon \mathbf{v}^{(l)}}\right) + \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) \\ &= d_{\mathcal{H}}\left(\mathbf{u}^{(l)} \odot \mathcal{K}_\varepsilon \mathbf{v}^{(l)}, \mathbf{a}\right) + \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) = d_{\mathcal{H}}\left(\mathbf{P}^{(l)}\mathbb{1}, \mathbf{a}\right) + \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}), \end{aligned}$$

de donde se deduce que

$$d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) \leq \frac{d_{\mathcal{H}}\left(\mathbf{P}^{(l)}\mathbb{1}, \mathbf{a}\right)}{1 - \lambda(\mathcal{K}_\varepsilon)^2}.$$

De forma análoga se deduce

$$\begin{aligned} d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) &\leq d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}^{(l+1)}) + d_{\mathcal{H}}(\mathbf{v}^{(l+1)}, \mathbf{v}) \leq d_{\mathcal{H}}\left(\mathbf{v}^{(l)}, \frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T \mathbf{u}^{(l+1)}}\right) + \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{u}^{(l)}, \mathbf{u}) \\ &= d_{\mathcal{H}}\left(\mathcal{K}_\varepsilon^T \mathbf{u}^{(l+1)} \odot \mathbf{v}^{(l)}, \mathbf{b}\right) + \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) = d_{\mathcal{H}}\left(\left(\mathbf{P}^{(l)}\right)^T \mathbb{1}, \mathbf{b}\right) + \lambda(\mathcal{K}_\varepsilon)^2\, d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) \end{aligned}$$

de donde se obtiene la desigualdad

$$d_{\mathcal{H}}(\mathbf{v}^{(l)}, \mathbf{v}) \leq \frac{d_{\mathcal{H}}\left(\left(\mathbf{P}^{(l)}\right)^T \mathbb{1}, \mathbf{b}\right)}{1 - \lambda(\mathcal{K}_\varepsilon)^2}.$$

□

En la siguiente proposición se proporciona la formulación dual del problema de transporte entrópico discreto. El problema primal (3.5) que hemos estudiado a lo largo del capítulo es un problema de minimización; sin embargo, el problema dual, el cual se va a introducir a continuación, es un problema de maximización.

Proposición 3.1.3 (Formulación dual del problema de transporte entrópico discreto)**.** *Sean* $\mathbf{a} \in \Sigma_m$ *y* $\mathbf{b} \in \Sigma_n$ *dos vectores de probabilidades y sea* $\mathbf{C} \in \mathbb{R}^{m\times n}$ *una matriz de coste. Sea* $\varepsilon > 0$ *el factor de regularización. El coste óptimo de transporte entrópico discreto es*

$$\mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a},\mathbf{b}) = \max_{\mathbf{f}\in\mathbb{R}^m,\,\mathbf{g}\in\mathbb{R}^n} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle - \varepsilon\left(\sum_{i,j} e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}}\right). \tag{3.16}$$

Además, los argumentos $\mathbf{f}$ *y* $\mathbf{g}$ *en los que se alcanza este máximo, a los cuales llamamos potenciales óptimos, satisfacen*

$$p_{i,j} = e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}}$$

donde $p_{i,j}$ *es la entrada* (i,j) *de la matriz* $\mathbf{P}^{\varepsilon}$ *de transporte entrópico. Decimos que esta es la formulación dual del problema de transporte entrópico discreto.*

Demostración. El Lagrangiano del problema de transporte entrópico discreto (3.5) se ha calculado en la Proposición 3.0.2:

$$\mathfrak{L}(\mathbf{P},\mathbf{f},\mathbf{g}) = \langle \mathbf{C},\mathbf{P}\rangle + \varepsilon\sum_{i,j} p_{i,j}(\log(p_{i,j})-1) - \sum_i f_i\cdot\left(\sum_j p_{i,j} - a_i\right) - \sum_j g_j\cdot\left(\sum_i p_{i,j} - b_j\right).$$

En esta proposición también se calculó que fijados $\mathbf{f}$ y $\mathbf{g}$, el mínimo de $\mathfrak{L}$ se alcanza en la matriz $\mathbf{P}^{\varepsilon} \in \mathcal{U}(\mathbf{a},\mathbf{b})$ cuyas entradas vienen dadas por $p_{i,j} = e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}}$. Por lo tanto, sustituyendo la expresión de $\mathbf{P}^{\varepsilon}$ en el Lagrangiano, obtenemos la siguiente función que depende de los parámetros $\mathbf{f}$ y $\mathbf{g}$:

$$\begin{aligned} g(\mathbf{f},\mathbf{g}) &= \min_{\mathbf{P}\in\mathbb{R}^{m\times n}} \mathfrak{L}(\mathbf{P},\mathbf{f},\mathbf{g}) \\ &= \sum_{i,j}(f_i+g_j-\varepsilon)\,e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}} + \sum_i f_i\,a_i - \sum_{i,j} f_i e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}} + \sum_j g_i\,b_j - \sum_{i,j} g_j e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}} \\ &= -\varepsilon\sum_{i,j} e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}} + \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle. \end{aligned}$$

Esta es la función dual de Lagrange definida en [11, Sec.5.1.2] a partir de la cual podemos plantear el problema dual de (3.5):

$$\max_{\mathbf{f}\in\mathbb{R}^m,\,\mathbf{g}\in\mathbb{R}^n} g(\mathbf{f},\mathbf{g}) = -\varepsilon\sum_{i,j} e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}} + \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle$$

Dado que la función objetivo del problema de transporte entrópico discreto es estrictamente convexa, las restricciones 3.1 son todas igualdades y existe al menos una solución factible (3.3) del problema, se satisfacen las condiciones de Slater para el problema primal 3.5. Estamos en las condiciones de aplicar el Principio de Dualidad fuerte (ver [11, Sec.5.2.3]), el cual nos asegura que el máximo de la función dual g coincide con el mínimo del problema primal (3.5) y por lo tanto se satisface (3.16). □

Observación 18. La caracterización que hemos dado de los potenciales $\mathbf{f},\mathbf{g}$ en los que se alcanza el máximo (3.16) en la formulación dual del problema de transporte entrópico discreto, nos permite relacionarlos con los vectores $\mathbf{u}$ y $\mathbf{v}$ definidos en (3.7) y que intervienen en el Algoritmo de Sinkhorn. Podemos ver que

$$\mathbf{u} = \log(\mathbf{f}) \quad \text{y} \quad \mathbf{v} = \log(\mathbf{g}),$$

donde entendemos que se realizan estas operaciones componente a componente. Esta relación nos permite deducir que los potenciales en los que se alcanza el óptimo son únicos salvo constante aditiva; es decir, el máximo (3.16) se alcanza en $\mathbf{f}', \mathbf{g}'$ si y solo si existe una constante K para la cual

$$\mathbf{f}' = \mathbf{f} + K \quad \text{y} \quad \mathbf{g}' = \mathbf{g} - K.$$

Este resultado se comprueba de forma sencilla a partir de la unicidad de los vectores $\mathbf{u}, \mathbf{v}$ que se ha probado en la Observación 14.

La relación que se ha descrito en la Observación 18, nos proporciona una forma de obtener los potenciales $(\mathbf{f}, \mathbf{g})$ en los que se alcanza el óptimo de (3.16). Para ello basta tomar el logaritmo de los vectores $\mathbf{u}$ y $\mathbf{v}$ que nos proporciona el Algoritmo de Sinkhorn. De hecho, la formulación dual del problema de transporte entrópico discreto proporciona otra perspectiva sobre el Algoritmo de Sinkhorn. Es posible probar que este algoritmo es equivalente a la obtención del máximo del problema dual mediante la maximización sucesiva de la función objetivo respecto de los vectores $\mathbf{f}$ y $\mathbf{g}$: Si consideramos un vector $\mathbf{g}^{(l)} \in \mathbb{R}^n$ fijado, podemos obtener el valor de $\mathbf{f}^{(l+1)} \in \mathbb{R}^m$ que maximize

$$\langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}^{(l)}, \mathbf{b} \rangle - \varepsilon \left(\sum_{i,j} e^{\frac{f_i + g_j^{(l)} - c_{i,j}}{\varepsilon}} \right). \tag{3.17}$$

El gradiente respecto de $\mathbf{f}$ de la función objetivo (3.17) viene dado por

$$\mathbf{a} - e^{\mathbf{f}/\varepsilon} \odot \left(\sum_{j} e^{\frac{f_i + g_j^{(l)} - c_{i,j}}{\varepsilon}} \right) = \mathbf{a} - e^{\mathbf{f}/\varepsilon} \odot \left(\mathcal{K}_\varepsilon e^{\mathbf{g}^{(l)}/\varepsilon} \right).$$

Igualando el gradiente anterior a 0 se deduce que el vector en el que se alcanza el óptimo es $\mathbf{f}^{(l+1)} = \varepsilon \log \left(\frac{\mathbf{a}}{\mathcal{K}_\varepsilon e^{\mathbf{g}^{(l)}/\varepsilon}} \right)$ por la condición necesaria de extremo relativo. Fijando ahora el vector $\mathbf{f}^{(l+1)}$ que se ha calculado, se puede maximizar (3.16) respecto de $\mathbf{g}$ de una forma similar con el fin de obtener el iterante $\mathbf{g}^{(l+1)}$. En este caso el gradiente de la función objetivo es

$$\mathbf{b} - e^{\mathbf{g}/\varepsilon} \odot \left(\sum_{i} e^{\frac{f_i^{(l+1)} + g_j - c_{i,j}}{\varepsilon}} \right) = \mathbf{b} - e^{\mathbf{g}/\varepsilon} \odot \left(\mathcal{K}_\varepsilon^T e^{\mathbf{f}^{(l+1)}/\varepsilon} \right)$$

de donde se deduce que en $\mathbf{g}^{(l+1)} = \varepsilon \log \left(\frac{\mathbf{b}}{\mathcal{K}_\varepsilon^T e^{\mathbf{f}^{(l+1)}/\varepsilon}} \right)$ se alcanza el máximo. Definiendo los vectores $\mathbf{u}^{(l)} = e^{\mathbf{f}^{(l)}/\varepsilon} \in \mathbb{R}^m_{>0}$ y $\mathbf{v}^{(l)} = e^{\mathbf{g}^{(l)}/\varepsilon} \in \mathbb{R}^n_{>0}$ obtenemos las iteraciones del Algoritmo de Sinkhorn.

3.2. Transporte entrópico discreto generalizado

Es posible dar una formulación más general del problema de transporte entrópico discreto. Fijada una matriz de coste $\mathbf{C} \in \mathbb{R}^{m \times n}$ y un factor de regularización $\varepsilon > 0$, el problema de transporte entrópico discreto generalizado viene dado por el problema de minimización

$$\min_{\substack{\mathbf{P}\mathbb{1}_n = \mathbf{a},\, \mathbf{P}^T\mathbb{1}_m = \mathbf{b} \\ \mathbf{P} \in \mathbb{R}^{m \times n}_{>0}}} \langle \mathbf{P}, C \rangle + \varepsilon \langle \mathbf{P}, \log(\mathbf{P}) - \mathbb{1} \rangle + F(\mathbf{a}) + G(\mathbf{b}) \tag{3.18}$$

donde F y G son funciones convexas definidas en $\mathbb{R}^m$ y $\mathbb{R}^n$ respectivamente. La hipótesis de convexidad de F y G implica la convexidad estricta de la función objetivo, la cual asegura la unicidad de la matriz minimizadora en

caso de existir. Para ver la relación entre el problema (3.18) y (3.5) consideramos dos vectores de probabilidades $\mathbf{a} \in \Sigma_m$ y $\mathbf{b} \in \Sigma_n$ fijados. Si tomamos las funciones

$$F(x) = \begin{cases} 0 & \text{si } x = \mathbf{a} \\ \infty & \text{si } x \neq \mathbf{a} \end{cases} \quad \text{y} \quad G(y) = \begin{cases} 0 & \text{si } y = \mathbf{b} \\ \infty & \text{si } y \neq \mathbf{b}, \end{cases}$$

las cuales son convexas, recuperamos el problema de transporte entrópico discreto (3.5). Un cálculo similar al que se ha realizado en la prueba de la Proposición 3.0.2 permiten comprobar que la función Lagrangiana asociada al problema de minimización (3.18) es

$$\begin{aligned} \mathcal{L}(\mathbf{P}, \mathbf{a}, \mathbf{b}, \mathbf{f}, \mathbf{g}) &= \langle \mathbf{P}, C \rangle + \varepsilon \langle \mathbf{P}, \log(\mathbf{P}) - \mathbb{1} \rangle + F(\mathbf{a}) + G(\mathbf{b}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1} - \mathbf{a} \rangle - \langle \mathbf{g}, \mathbf{P}^T\mathbb{1} - \mathbf{b} \rangle \\ &= \langle \mathbf{P}, C \rangle + \varepsilon \langle \mathbf{P}, \log(\mathbf{P}) - \mathbb{1} \rangle - \langle \mathbf{f}, \mathbf{P}\mathbb{1} \rangle - \langle \mathbf{g}, \mathbf{P}^T\mathbb{1} \rangle \\ &\quad - (\langle -\mathbf{f}, \mathbf{a} \rangle - F(\mathbf{a})) - (\langle -\mathbf{g}, \mathbf{b} \rangle - G(\mathbf{b})). \end{aligned}$$

La derivada parcial de la función $\mathcal{L}$ respecto de $p_{i,j}$ es de la forma

$$\frac{\partial \mathcal{L}}{\partial p_{i,j}}(\mathbf{P}, \mathbf{a}, \mathbf{b}, \mathbf{f}, \mathbf{g}) = c_{i,j} + \varepsilon \log(p_{i,j}) - f_i - g_j.$$

Igualando a cero la expresión anterior y despejando $p_{i,j}$ se deduce que la matriz $\mathbf{P}^{\varepsilon}_{F,G}$ en la que se alcance el óptimo del problema de transporte entrópico discreto generalizado debe admitir una descomposición de la forma

$$\mathbf{P}^{\varepsilon}_{F,G} = diag(\mathbf{u}) \cdot \mathcal{K}_{\varepsilon} \cdot diag(\mathbf{v}),$$

con $\mathbf{u} \in \mathbb{R}^m_{>0}$ y $\mathbf{v} \in \mathbb{R}^n_{>0}$. Estos vectores vienen dados por $\mathbf{u} = e^{\mathbf{f}/\varepsilon}$ y $\mathbf{v} = e^{\mathbf{g}/\varepsilon}$.

En la siguiente definición se introduce la transformada de Legendre, la cual va a intervenir en la formulación dual del problema de transporte entrópico generalizado.

Definición 3.2.1. Sea f una función de $\mathbb{R}^n$. La función de $\mathbb{R}^n$

$$f^*(p) = \sup_{x \in \mathbb{R}^n} \langle x, p \rangle - f(x)$$

es la transformada de Legendre de f.

Se puede comprobar que fijados $\mathbf{P}$ y los valores de los multiplicadores $\mathbf{f}, \mathbf{g}$, los óptimos $\mathbf{a}^*$ y $\mathbf{b}^*$ que minimizan el Lagrangiano verifican:

$$\begin{aligned} F^*(-\mathbf{f}) &= \sup_{\mathbf{a}} \langle -\mathbf{f}, \mathbf{a} \rangle - F(\mathbf{a}) = \langle -\mathbf{f}, \mathbf{a}^* \rangle - F(\mathbf{a}^*) \\ G^*(-\mathbf{g}) &= \sup_{\mathbf{b}} \langle -\mathbf{g}, \mathbf{b} \rangle - G(\mathbf{b}) = \langle -\mathbf{g}, \mathbf{b}^* \rangle - G(\mathbf{b}^*). \end{aligned} \tag{3.19}$$

El función objetivo del problema dual de (3.18) se obtiene minimizando $\mathcal{L}$ respecto de $\mathbf{P}, \mathbf{a}, \mathbf{b}$. Esto es equivalente a sustituir en $\mathcal{L}$ la expresión para la matriz óptima $\mathbf{P}^{\varepsilon}_{F,G}$ que se ha calculado y las expresiones (3.19). Entonces, la función objetivo del problema dual es

$$\begin{aligned} g(\mathbf{f}, \mathbf{g}) = \min_{\mathbf{P}, \mathbf{a}, \mathbf{b}} \mathcal{L}(\mathbf{P}, \mathbf{a}, \mathbf{b}, \mathbf{f}, \mathbf{g}) &= \sum_{i,j} c_{i,j} e^{\frac{f_i + g_j - c_{i,j}}{\varepsilon}} + \varepsilon \sum_{i,j} \left(\frac{f_i + g_j - c_{i,j}}{\varepsilon} - 1 \right) e^{\frac{f_i + g_j - c_{i,j}}{\varepsilon}} \\ &\quad - \sum_{i,j} f_i e^{\frac{f_i + g_j - c_{i,j}}{\varepsilon}} - \sum_{i,j} g_j e^{\frac{f_i + g_j - c_{i,j}}{\varepsilon}} - F^*(-\mathbf{f}) - G^*(-\mathbf{g}) \\ &= -F^*(-\mathbf{f}) - G^*(-\mathbf{g}) - \varepsilon \sum_{i,j} e^{\frac{f_i + g_j - c_{i,j}}{\varepsilon}} \end{aligned} \tag{3.20}$$

y por lo tanto

$$\max_{\mathbf{f}\in\mathbb{R}^m,\mathbf{g}\in\mathbb{R}^n} -F^*(-\mathbf{f}) - G^*(-\mathbf{g}) - \varepsilon \sum_{i,j} e^{\frac{f_i+g_j-c_{i,j}}{\varepsilon}} \tag{3.21}$$

es la formulación dual del problema de transporte entrópico discreto generalizado. Por el Principio de Dualidad fuerte, el cual se aplica por tenerse la condiciones de Slater, los óptimos del problema primal y el dual coinciden.

La formulación dual que se ha obtenido permite establecer un método para el cálculo de la solución del problema de transporte entrópico discreto generalizado. Empleando una técnica similar a la adoptada en el problema (3.5) se puede implementar un algoritmo para obtener el máximo (3.21): en cada iteración se fija el parámetro $\mathbf{g}$ y se busca el parámetro $\mathbf{f}$ en el que se alcance el óptimo. Seguidamente se considera fijado el parámetro $\mathbf{f}$ que se acaba de calcular y se maximiza (3.21) respecto del vector $\mathbf{g}$ obteniendo un nuevo iterarte. Este esquema se denomina Algoritmo de Sinkhorn generalizado. En el Algoritmo 2 se muestra un pseudocódigo para la implementación de este.

Algoritmo 2: Algoritmo de Sinkhorn generalizado

Input: $\mathbf{C}, F, G, \varepsilon, N$
Output: $\mathbf{u}, \mathbf{v}, \mathbf{P}$
$\mathbf{g}^{(0)} \leftarrow \mathbf{0}$;
while $l \leq N-1$ **do**
  $\mathbf{f}^{(l+1)} \leftarrow \max_{\mathbf{f}\in\mathbb{R}^m} -F^*(-\mathbf{f}) - G^*(-\mathbf{g}^{(l)}) - \varepsilon \sum_{i,j} e^{\frac{f_i+g_j^{(l)}-c_i,j}{\varepsilon}}$;
  $\mathbf{g}^{(l+1)} \leftarrow \max_{\mathbf{g}\in\mathbb{R}^n} -F^*(-\mathbf{f}^{(l+1)}) - G^*(-\mathbf{g}) - \varepsilon \sum_{i,j} e^{\frac{f_i^{(l+1)}+g_j-c_i,j}{\varepsilon}}$;
end
$\mathbf{u} \leftarrow \mathbf{u}^{(N)}$;
$\mathbf{v} \leftarrow \mathbf{v}^{(N)}$;
$\mathbf{P} \leftarrow diag(\mathbf{u}) \cdot \mathcal{K}_\varepsilon \cdot diag(\mathbf{v})$

3.3. Diferenciabilidad respecto de los argumentos

A continuación se va a estudiar una serie de resultados que tienen como fin probar la diferenciabilidad del coste de transporte entrópico discreto respecto $\mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ respecto de cada uno de sus argumentos: los vectores de probabilidades $\mathbf{a}$, $\mathbf{b}$ y la matriz de coste $\mathbf{C}$. Esta propiedad es muy relevante para el estudio de problemas del cálculo variacional en los que se emplee el coste de transporte entrópico como función de pérdida. Se profundizará en estas aplicaciones en el Capítulo 4.

En primer lugar, vamos a introducir un resultado que nos proporciona el gradiente de $\mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ respecto de los vectores de probabilidades $\mathbf{a}$ y $\mathbf{b}$.

Observación 19. Debemos tener en cuenta que el conjunto $\Sigma_m \times \Sigma_n$ donde está definida la función $(\mathbf{a},\mathbf{b}) \mapsto \mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ no es un conjunto abierto de $\mathbb{R}^m \times \mathbb{R}^n$. Por ello, no tiene sentido considerar la diferencial de esta aplicación. Sin embargo, $\Sigma_m \times \Sigma_n$ es una variedad diferenciable cuyo espacio tangente en un punto $(\mathbf{a},\mathbf{b})$ está dado por el producto de los espacios tangentes de Σ_m en $\mathbf{a}$ y Σ_n en $\mathbf{b}$. Es fácil comprobar que Σ_m es una variedad de dimensión $m-1$ y que su espacio tangente en un punto se puede identificar de forma natural con el subespacio $m-1$ dimensional de $\mathbb{R}^m$ dado por los vectores cuyas componentes suman 0.

El resultado que se va a presentar estudia la diferenciabilidad de $(\mathbf{a},\mathbf{b}) \mapsto \mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a},\mathbf{b})$ como función definida en una variedad diferenciable. En este contexto, sí tiene sentido considerar su diferencial en un punto, la cual es una aplicación lineal que actúa sobre el espacio tangente en ese punto. Por la observación que hemos

dado, podemos identificar el espacio tangente en un punto de $\Sigma_m \times \Sigma_n$ con el conjunto de pares de vectores $(\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^m \times \mathbb{R}^n$ tales que la suma de las componentes de $\mathbf{w}_1$ es 0 y la suma de las componentes de $\mathbf{w}_2$ es 0. Entonces, cada aplicación lineal sobre el espacio tangente en un punto de $\Sigma_m \times \Sigma_n$ se puede representar con un vector $(\mathbf{u}_1, \mathbf{u}_2) \in \mathbb{R}^m \times \mathbb{R}^n$ de forma que la evaluación de esta en $(\mathbf{w}_1, \mathbf{w}_2)$ sea

$$\langle (\mathbf{u}_1, \mathbf{u}_2), (\mathbf{w}_1, \mathbf{w}_2) \rangle.$$

Esta representación no es única porque para cada $K_1, K_2 \in \mathbb{R}$ el vector $(\mathbf{u}_1 + K_1, \mathbf{u}_2 + K_2)$ también verifica la afirmación anterior:

$$\langle (\mathbf{u}_1 + K_1, \mathbf{u}_2 + K_2), (\mathbf{w}_1, \mathbf{w}_2) \rangle = \langle (\mathbf{u}_1, \mathbf{u}_2), (\mathbf{w}_1, \mathbf{w}_2) \rangle + \langle (K_1 \mathbb{1}, K_2 \mathbb{1}), (\mathbf{w}_1, \mathbf{w}_2) \rangle = \langle (\mathbf{u}_1, \mathbf{u}_2), (\mathbf{w}_1, \mathbf{w}_2) \rangle.$$

La última igualdad se debe a que $\langle (K_1 \mathbb{1}, K_2 \mathbb{1}), (\mathbf{w}_1, \mathbf{w}_2) \rangle = K_1 \langle \mathbb{1}, \mathbf{w}_1 \rangle + K_2 \langle \mathbb{1}, \mathbf{w}_2 \rangle = K_1 \cdot 0 + K_2 \cdot 0 = 0$.

La Observación 19 nos indica que el gradiente de $(\mathbf{a}, \mathbf{b}) \mapsto \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ es un vector de $\mathbb{R}^m \times \mathbb{R}^n$ que está definido salvo constantes en el sentido que se ha precisado.

Proposición 3.3.1. *Sea $\mathbf{C}$ una matriz de coste de orden $m \times n$ y sea $\varepsilon < 0$ un factor de regularización fijado. Consideramos la función $(\mathbf{a}, \mathbf{b}) \mapsto \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ definida en $\Sigma_m \times \Sigma_n$. Esta función es continuamente diferenciable y su gradiente es*

$$\nabla_{\mathbf{a},\mathbf{b}} \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = (\mathbf{f}, \mathbf{g})$$

donde $(\mathbf{f}, \mathbf{g})$ son unos potenciales en los que se alcanza el óptimo del problema dual del transporte entrópico discreto.

Demostración. Podemos ver que esta afirmación es coherente con la Observación 19 ya que cualquier par de potenciales óptimos es de la forma $(\mathbf{f} + K, \mathbf{g} - K)$ para una constante K. La demostración del resultado se puede encontrar en [12, Prop.2.3.]. □

Ahora nos dedicaremos al estudio de la diferenciabilidad del coste de transporte entrópico respecto de la matriz de coste. Para ello, vamos a emplear un teorema que enunciamos a continuación.

Teorema 3.3.1. *Sea Θ un abierto conexo de $\mathbb{R}^d$ y sean $\mathbf{a} : \theta \mapsto \mathbf{a}(\theta) \in \mathbb{R}^m_>$, $\mathbf{b} : \theta \mapsto \mathbf{b}(\theta) \in R^n_>$, $\mathbf{C} : \theta \mapsto \mathbf{C}(\theta) \in \mathbb{R}^{m \times n}$ y $\varepsilon : \theta \mapsto \varepsilon(\theta) > 0$ aplicaciones de clase $\mathcal{C}^2$ en Θ. Entonces, la aplicación*

$$\mathbf{P}^{\varepsilon} : \theta \mapsto \mathbf{P}^{\varepsilon}(\theta)$$

que asocia a cada parámetro θ la solución del problema de transporte entrópico discreto entre $\mathbf{a}(\theta)$ y $\mathbf{b}(\theta)$ con matriz de coste $\mathbf{C}(\theta)$ y factor de regularización $\varepsilon(\theta)$, es continuamente diferenciable en Ω. Además, si para cada $k \geq 1$ se considera la aplicación que asocia cada parámetro θ con el iterante k-ésimo de Algoritmo de Sinkhorn, $\mathbf{P}^{(k)} : \theta \mapsto \mathbf{P}^{(k)}(\theta)$, se tiene que esta aplicación es continuamente diferenciable y

$$\lim_{k \to \infty} \frac{d\mathbf{P}^{(k)}}{d\theta}(\theta) = \frac{d\mathbf{P}^{\varepsilon}}{d\theta}(\theta).$$

Demostración. La demostración del teorema se puede encontrar en el Teorema 3.3 de [13]. □

Corolario 3.3.1. *La matriz en la que se alcanza el óptimo del problema de transporte entrópico discreto, $\mathbf{P}^{\varepsilon}$, es continuamente diferenciable respecto de todos sus argumentos: los vectores de probabilidades $\mathbf{a}$, $\mathbf{b}$, la matriz de coste $\mathbf{C}$ y el factor de regularización ε.*

Demostración. Vamos a probar que la solución del problema del transporte entrópico discreto $\mathbf{P}^\varepsilon$ es continuamente diferenciable respecto de $\mathbf{C}$ si se fijan el resto de argumentos, es decir, los vectores $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^n$ y ε. Basta considerar el abierto $\Theta = \mathbb{R}^{m\times n}$, la parametrización de la función de coste dada por la identidad

$$\mathbf{C} : \theta \mapsto \mathbf{C}(\theta) = \theta$$

y las parametrizaciones constantes de $\mathbf{a}$, $\mathbf{b}$ y ε: $\mathbf{a}(\theta) = \mathbf{a},\ \mathbf{b}(\theta) = \mathbf{b},\ \varepsilon(\theta) = \varepsilon$, para cada $\theta \in \Theta$. Todas estas aplicaciones son de clase $\mathcal{C}^2$ en Θ por lo que se aplica el Teorema 3.3.1 de donde se concluye el resultado.

La demostración de la diferenciabilidad respecto del resto de parámetros es análoga a la que hemos desarrollado para el caso de $\mathbf{C}$. □

Los resultados previos nos permiten obtener una expresión para el gradiente de $\mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ respecto de la matriz de coste $\mathbf{C}$. En la siguiente proposición se recoge esta expresión que intervendrá en resultados del Capítulo 4.

Proposición 3.3.2. *Sean $\mathbf{a} \in \Sigma_m$ y $\mathbf{b} \in \Sigma_n$ dos vectores de probabilidad. Sea $\varepsilon > 0$ un factor de regularización. Consideramos la función*

$$\mathbf{C} \mapsto \mathcal{R}(\mathbf{C}) \overset{def.}{=} \mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$$

definida en el conjunto de matrices de orden $m \times n$. Esta función es diferenciable y su gradiente es

$$\nabla\mathcal{R}(\mathbf{C}) = \mathbf{P}^\varepsilon$$

donde $\mathbf{P}^\varepsilon$ es la matriz asociada a la solución del problema de transporte entrópico discreto con matriz de coste $\mathbf{C}$. Es decir, se verifica $\frac{\partial\mathcal{R}}{\partial c_{i,j}}(\mathbf{C}) = p^\varepsilon_{i,j}$ para cada $1 \le i \le m$ y $1 \le j \le n$.

Demostración. Dada una matriz de coste $\mathbf{C} \in \mathbb{R}^{m\times n}$ denotamos por $\mathbf{P}^\varepsilon(\mathbf{C})$ a la única solución del problema de transporte entrópico discreto entre $\mathbf{a}$ y $\mathbf{b}$ con matriz de coste $\mathbf{C}$. Es decir, $\mathbf{P}^\varepsilon(\mathbf{C})$ a la única matriz de $\mathcal{U}(\mathbf{a}, \mathbf{b})$ que satisface

$$\mathcal{R}(\mathbf{C}) = \mathcal{L}^\varepsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) = \langle \mathbf{P}^\varepsilon(\mathbf{C}), \mathbf{C}\rangle + \varepsilon \langle \mathbf{P}^\varepsilon(\mathbf{C}), \log(\mathbf{P}^\varepsilon(\mathbf{C})) - \mathbb{1}\rangle.$$

Por el Corolario 3.3.1 podemos asegurar que $\mathbf{P}^\varepsilon(\mathbf{C})$ es una aplicación continuamente diferenciable respecto de $\mathbf{C}$ y por lo tanto, la función $\mathcal{R}$ también lo es por ser composición de aplicaciones diferenciables. Vamos a calcular las derivadas parciales de $\mathcal{R}$ respecto de cada componente de $\mathbf{C}$:

$$\begin{aligned}
\frac{\partial\mathcal{R}}{\partial c_{i,j}}(\mathbf{C}) &= \sum_{k,l}\left(\frac{\partial p^\varepsilon_{k,l}}{\partial c_{i,j}}(\mathbf{C})\cdot c_{k,l} + p^\varepsilon_{k,l}(\mathbf{C})\cdot\frac{\partial c_{k,l}}{\partial c_{i,j}}(\mathbf{C})\right)\\
&\quad + \varepsilon\sum_{k,l}\left(\frac{\partial p^\varepsilon_{k,l}}{\partial c_{i,j}}(\mathbf{C})\cdot(\log(p^\varepsilon_{k,l}(\mathbf{C})) - 1) + p^\varepsilon_{k,l}(\mathbf{C})\cdot\frac{1}{p^\varepsilon_{k,l}(\mathbf{C})}\cdot\frac{\partial p^\varepsilon_{k,l}}{\partial c_{i,j}}(\mathbf{C})\right)\\
&= p^\varepsilon_{i,j}(\mathbf{C}) + \sum_{k,l}\left(\frac{\partial p^\varepsilon_{k,l}}{\partial c_{i,j}}(\mathbf{C})\cdot c_{k,l}\right) + \varepsilon\sum_{k,l}\left(\frac{\partial p^\varepsilon_{k,l}}{\partial c_{i,j}}(\mathbf{C})\cdot\log(p^\varepsilon_{k,l}(\mathbf{C}))\right)
\end{aligned}$$

Desarrollando el término $\log(p^\varepsilon_{k,l}(\mathbf{C}))$ atendiendo a la estructura de la matriz $\mathbf{P}^\varepsilon(\mathbf{C})$ obtenemos

$$\log(p^\varepsilon_{k,l}(\mathbf{C})) = \log\left(u_k(\mathbf{C})\cdot e^{-\frac{c_{k,l}}{\varepsilon}}\cdot v_l(\mathbf{C})\right) = \log(u_k(\mathbf{C})) - \frac{c_{k,l}}{\varepsilon} + \log(v_l(\mathbf{C})).$$

En la expresión previa intervienen las componentes de los vectores $\mathbf{u}$ y $\mathbf{v}$ que hemos introducido en (3.7). A partir de esta igualdad podemos escribir

$$\begin{aligned}\frac{\partial \mathcal{R}}{\partial c_{i,j}}(\mathbf{C}) &= p^{\varepsilon}_{i,j}(\mathbf{C}) + \sum_{k,l}\left(\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C})\cdot c_{k,l}\right) + \varepsilon\sum_{k,l}\left(\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C})\cdot \log(p^{\varepsilon}_{k,l}(\mathbf{C}))\right)\\ &= p^{\varepsilon}_{i,j}(\mathbf{C}) + \varepsilon\sum_{k,l}\left(\log(u_k(\mathbf{C}))\cdot\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C})\right) + \varepsilon\sum_{k,l}\left(\log(v_l(\mathbf{C}))\cdot\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C})\right)\\ &= p^{\varepsilon}_{i,j}(\mathbf{C}) + \varepsilon\sum_{k}\left(\log(u_k(\mathbf{C}))\cdot\sum_{l}\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C})\right) + \varepsilon\sum_{l}\left(\log(v_l(\mathbf{C}))\cdot\sum_{k}\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C})\right)\end{aligned}$$

Para concluir la demostración, basta aplicar la linealidad de la derivada de donde se deduce

$$\sum_{l}\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C}) = \frac{\partial\left(\sum_l p^{\varepsilon}_{k,l}\right)}{\partial c_{i,j}}(\mathbf{C}) = \frac{\partial \mathbf{a}}{\partial c_{i,j}}(\mathbf{C}) = 0$$

$$\sum_{k}\frac{\partial p^{\varepsilon}_{k,l}}{\partial c_{i,j}}(\mathbf{C}) = \frac{\partial\left(\sum_k p^{\varepsilon}_{k,l}\right)}{\partial c_{i,j}}(\mathbf{C}) = \frac{\partial \mathbf{b}}{\partial c_{i,j}}(\mathbf{C}) = 0.$$

En consecuencia, se tiene $\frac{\partial \mathcal{R}}{\partial c_{i,j}}(\mathbf{C}) = p^{\varepsilon}_{i,j}(\mathbf{C})$. □

Capítulo 4

Problemas variacionales

Este ultimo capítulo está dedicado al estudio de varios problemas de optimización asociados al transporte entrópico. En la mayoría de situaciones prácticas en las que aparecen estos problemas se tienen distribuciones continuas con las cuales no es posible trabajar de forma computacional. Por ello, se utilizan técnicas de discretización que permiten simplificar el problema y realizar cálculos con distribuciones discretas. De esta forma, se puede transformar un problema variacional inicial en otro en el cual se pueden emplear las herramientas que se han desarrollado en el Capitulo 3. Se van a introducir dos tipos de discretizaciones: Euleriana y Lagrangiana. Se describirá como tratar con ellas y qué ventajas proporcionan cada una de ellas.

Los dos primeros problemas que se van a estudiar se pueden enmarcar como un problema de minimización

$$\min_{\alpha \in \mathcal{S}} \mathcal{F}(\alpha) \tag{4.1}$$

de un funcional $\mathcal{F}$ definido en un subconjunto $\mathcal{S}$ de $\mathcal{P}(\mathcal{X})$. Se va a considerar que funcionales que en los cuales intervenga un coste de transporte óptimo respecto de ciertas probabilidades fijadas. En los casos que vamos a tratar los funcionales que se busca minimizar van a depender de la distancia p de Wasserstein a unas probabilidades de referencia por lo que se restringirá la búsqueda del mínimo a un subconjunto $\mathcal{S}$ de $\mathcal{P}_p(\mathcal{X})$. De esta manera se va a poder garantizar que $\mathcal{F}$ es finito en $\mathcal{S}$. El problema (4.1) aparece en muchos ejemplos del área de aprendizaje automático, donde el papel del funcional $\mathcal{F}$ es el de función de pérdida. En esta situación se busca obtener el mínimo de esta función y con más frecuencia, un elemento minimizador de $\mathcal{F}$. En modelos de aprendizaje automático, los elementos del conjunto factible $\mathcal{S}$ suelen estar parametrizados por una aplicación definida en un subconjunto Θ de $\mathbb{R}^d$, de forma que el problema (4.1) se puede entender como la búsqueda de los parámetros óptimos en los que se minimiza $\mathcal{F}$. Estos parámetros son los que determinan el modelo.

En la formulación general (4.1) no se puede afirmar la existencia del mínimo del funcional $\mathcal{F}$ ya que ni siquiera se puede asegurar que el conjunto $\{\mathcal{F}(\alpha) : \alpha \in \mathcal{S}\}$ esté acotado inferiormente. Por esta razón, en los dos problemas variacionales que se van a introducir se considerarán simplificaciones con las que se obtenga un funcional con la regularidad suficiente para el empleo de técnicas de minimización como el Algoritmo de Descenso de Gradiente. La principal suposición que se va a adoptar es que el conjunto $\mathcal{S}$ está formado por probabilidades con soporte finito. Esta hipótesis se considera por dos razones: en primer lugar, bajo esta condición se pueden emplear las técnicas del caso discreto que se han introducido en al Capítulo 3. El segundo argumento que sostiene esta simplificación se basa en la necesidad de empleo de técnicas de cálculo computacional para la resolución de estos problemas. Este hecho, obliga a almacenar las probabilidades continuas como un número finito de valores; es decir, se discretizan las probabilidades continuas. En la siguiente sección se va a detallar dos esquemas de discretización de probabilidades basados en dos enfoques distintos.

Otra simplificación que se va a introducir en el estudio de los problemas de la forma (4.1) consiste en sustituir la distancia p de Wasserstein, que interviene en el funcional $\mathcal{F}$, por la distancia de Wasserstein regularizada.

De este modo se obtiene un funcional alternativo $\tilde{\mathcal{F}}$ que aproxima al funcional original y además proporciona las ventajas del transporte entrópico frente al transporte óptico: en algunas situaciones se puede garantizar la convexidad estricta de $\tilde{\mathcal{F}}$, y además permite emplear el Algoritmo de Sinkhorn para su cálculo en el caso discreto. En consecuencia, el problema resultante se puede formular de la siguiente manera:

$$\min_{\alpha \in \mathcal{S}_F} \tilde{\mathcal{F}}(\alpha) \tag{4.2}$$

donde $\mathcal{S}_F$ es un subconjunto de $\mathcal{P}(\mathcal{X})$ formado por probabilidades con soporte finito.

4.1. Tipos de discretizaciones

Antes de introducir los problemas variacionales vamos a estudiar los dos tipos de discretizaciones que vamos a considerar. Según la discretización que se emplee, la formulación de cada problema será distinta. La situación que se plantea es la siguiente: se tiene una probabilidad continua P en $\mathbb{R}^d$ y se quiere trabajar con una probabilidad $\tilde{P}$ con soporte finito que concentre la masa en a los sumo m puntos de $\mathbb{R}^d$ y que se asemeje a la probabilidad original P.

4.1.1. Discretización Euleriana

La primera estrategia que planetamos para representar una probabilidad arbitraria P como una probabilidad con soporte finito se basa en fijar estos puntos de antemano. Para ello, se considera una malla finita de celdas que divida el soporte de P, el cual supondremos que es compacto. En cada una de las celdas seleccionadas se elige un punto que representará a toda la celda al cual llamaremos centro. De esta forma, el soporte de la probabilidad discretizada estará contenido en los centros de esta malla y vendrá determinada por la probabilidad de la celda correspondiente.

A continuación se van a introducir el concepto de partición centrada en un compacto, el cual formaliza la noción de malla que hemos empleado. A partir, de este concepto vamos a definir la discretización Euleriana.

Definición 4.1.1. Sea K un conjunto compacto de $\mathbb{R}^d$. Una partición centrada de K es un conjunto finito de pares $\mathcal{T} = \{(T_k, x_k)\}_{k=1}^m$ que satisfacen:

1. T_k es un subconjunto medible de K para cada $1 \leq k \leq m$. Diremos que el conjunto T_k es una celda de $\mathcal{T}$.
2. $T = \{T_k\}_{k=1}^m$ es una partición de K.
3. Para cada k, $1 \leq k \leq m$ se tiene $x_k \in T_k$. Decimos que x_k es el centro de la celda T_k.

Escribimos $X = \{x_1, \ldots, x_m\}$ para denotar al conjunto de centros de $\mathcal{T}$.

Definición 4.1.2. Sea K un conjunto compacto de $\mathbb{R}^d$ y sea $\mathcal{T} = \{(T_k, x_k)\}_{k=1}^m$ una partición centrada de K. Definimos el diámetro de $\mathcal{T}$ como el diámetro de la partición T y lo denotamos

$$\delta(\mathcal{T}) \overset{\text{def.}}{=} \delta(T) = \max_{1 \leq k \leq m} \delta(T_k).$$

Definición 4.1.3. Sea P una probabilidad en $\mathbb{R}^m$ con soporte contenido en un conjunto compacto K y sea $\mathcal{T} = \{(T_k, x_k)\}_{k=1}^m$ una partición centrada de K. Denotamos por $P_{\mathcal{T}}$ a la probabilidad cuyo soporte es el conjunto de centros de $\mathcal{T}$ y función de masa de probabilidad

$$f_n(x_k) = P(T_k) \quad , \; x_k \in X.$$

Decimos que $P_{\mathcal{T}}$ es la discretización Euleriana de P en la partición centrada $\mathcal{T}$.

Se va a estudiar el comportamiento de asintótico de la dicretización Euleriana cuando la malla se hace más fina. Para que se pueda asegurar la convergencia de la discretización a la probabilidad original P, se debe imponer ciertas condiciones a la malla: es necesario que la malla se vaya "refinando" en un sentido que vamos a definir con el concepto de sucesión admisible de de particiones centradas de un compacto.

Definición 4.1.4. Sea K un conjunto compacto de $\mathbb{R}^d$. Sea $\{\mathcal{T}_n\}_{n\geq 1}$ una sucesión de particiones centradas de K

$$\mathcal{T}_n = \{(T_{k,n}, x_{k,n})\}_{k=1}^{m_n}.$$

Decimos que la sucesión $\{\mathcal{T}_n\}_{n\geq 1}$ es admisible si satisface:

1. La partición $T_{n+1} = \{T_{k,n+1}\}_{k=1}^{m_{n+1}}$ es un refinamiento de la partición $T_n = \{T_{k,n}\}_{k=1}^{m_n}$.
2. Sean $X_n = \{x_{k,n}\}_{k=1}^{m_n}$ y $X_{n+1} = \{x_{k,n+1}\}_{k=1}^{m_{n+1}}$ los conjuntos de centros de $\mathcal{T}_n$ y $\mathcal{T}_{n+1}$ respectivamente. Entonces $X_n \subset X_{n+1}$.
3. El diámetro de las particiones T_n tiende a 0:

$$\delta(T_n) = \max_{1\leq k\leq m_n} \delta(T_{k,n}) \xrightarrow[n\to\infty]{} 0.$$

Observación 20. Las condiciones 1 y 2 de la definición 4.1.4 son equivalentes a que todo conjunto $T_{k,n}$ de T_n se puede escribir como una unión finita

$$\bigcup_{s\in F} T_{s,n+1}$$

tal que $x_{k,n}$ es el centro de $T_{s,n+1}$ para algún $s \in F$.

Lema 4.1.1. Sea K un subconjunto compacto de $\mathbb{R}^d$ y sea $\{\mathcal{T}_n\}_{n\geq 1}$ una sucesión admisible de particiones centradas de K. Sea $B(x,r)$ una bola abierta contenida en K. Entonces, existe un $n \geq 1$ y $1 \leq k \leq m_n$ tal que $x \in T_{k,n} \subset B(x,r)$.

Demostración. Por ser $\{\mathcal{T}_n\}_{n\geq 1}$ una sucesión admisible, se deduce que existe $n \geq 1$ tal que $\delta(T_n) < r$. Además, por ser T_n una partición, se deduce que existe $T_{k,n}$ tal que $x \in T_{k,n}$. Entonces, para cada $y \in T_{k,n}$ se verifica

$$\|x - y\|_2 \leq \sup_{y_1,y_2\in T_{k,n}} \|y_1 - y_2\|_2 = \delta(T_{k,n}) \leq \delta(T_n) < r$$

y en consecuencia $T_{k,n} \subset B(x,r)$. □

Teorema 4.1.2. *Sea P una probabilidad de $\mathbb{R}^d$ con soporte contenido en un compacto K y sea $\{\mathcal{T}\}_{n\geq 1}$ una sucesión admisible de particiones centradas de K. Entonces se tiene la convergencia débil*

$$P_{\mathcal{T}_n} \xrightarrow{w} P.$$

Demostración. Sea $A \subset \mathbb{R}^m$ un conjunto medible con frontera de probabilidad nula, es decir, $P(Fr(A)) = 0$. Para cada $n \geq 1$ consideramos los siguientes conjuntos

$$\text{INT}_n(A) = \bigcup_{T_{k,n}\subset A} T_{k,n} \quad \text{y} \quad \text{SUP}_n(A) = \bigcup_{T_{k,n}\cap A\neq\emptyset} T_{k,n}$$

los cuales satisfacen

$$\text{INT}_n(A) \subset A \subset \text{SUP}_n(A). \tag{4.3}$$

Por la relación (4.3), se deducen las siguientes desigualdades:

$$P_{\mathcal{T}_n}(\text{INT}_n(A)) \leq P_{\mathcal{T}_n}(A) \leq P_{\mathcal{T}_n}(\text{SUP}_n(A))$$

$$P(\text{INT}_n(A)) \leq P(A) \leq P(\text{SUP}_n(A)).$$

Además, se dan las siguientes igualdades

$$P_{\mathcal{T}_n}(\text{INT}_n(A)) = \sum_{x_k \in \text{INT}_n(A)} f_n(x_k) = \sum_{T_{k,n} \subset A} P(T_{k,n}) = P\left(\bigsqcup_{T_{k,n} \subset A} T_{k,n} \right) = P(\text{INT}_n(A))$$

$$P_{\mathcal{T}_n}(\text{SUP}_n(A)) = \sum_{x_k \in \text{SUP}_n(A)} f_n(x_k) = \sum_{T_{k,n} \cap A \neq \emptyset} P(T_{k,n}) = P\left(\bigsqcup_{T_{k,n} \cap A \neq \emptyset} T_{k,n} \right) = P(\text{SUP}_n(A)).$$

Vamos a probar dos inclusiones que van a ser clave para probar el resultado:

$$\text{INT}_n(A) \subset \text{INT}_{n+1}(A) \tag{4.4}$$

$$\text{SUP}_{n+1}(A) \subset \text{SUP}_n(A) \tag{4.5}$$

Para probar la expresión (4.4) tomamos $x \in \text{INT}_n(A)$, entonces $x \in T_{k,n}$ para algún $1 \leq k \leq m_n$. Por definición de $\text{INT}_n(A)$ se verifica $T_{k,n} \subset A$. Entonces existe $1 \leq k' \leq m_{n+1}$ tal que $x \in T_{k',m+1}$. Dado que T_{n+1} es un refinamiento de T_n, se tiene $T_{k',n+1} \subset T_{k,n}$ y por lo tanto $T_{k',n+1} \subset A$. En consecuencia, $T_{k',n+1} \subset \text{INT}_{n+1}(A)$ y $x \in \text{INT}_{n+1}(A)$.

Para probar la inclusión(4.5) seguimos un argumento similar al anterior: tomamos $x \in \text{SUP}_{n+1}(A)$, entonces $x \in T_{k,n+1}$ para algún $1 \leq k \leq m_{n+1}$. Por definición de $\text{SUP}_{n+1}(A)$ se verifica $T_{k,n+1} \cap A \neq \emptyset$. Entonces existe $1 \leq k' \leq m_n$ tal que $x \in T_{k',n}$. Dado que T_{n+1} es un refinamiento de T_n, se verifica $T_{k,n+1} \subset T_{k',n}$ y por lo tanto $T_{k',n} \cap A \supset T_{k,n+1} \cap A \neq \emptyset$. En consecuencia, $T_{k',n} \subset \text{SUP}_n(A)$ y $x \in \text{SUP}_n(A)$.

Las inclusiones (4.4) y (4.5) nos permiten asegurar que existen los límites de los conjuntos $\text{INT}_n(A)$ y $\text{SUP}_n(A)$ respectivamente. Vamos a demostrar la siguiente cadena de inclusiones:

$$\mathring{A} \subset \lim_{m\to\infty} \text{INT}_n(A) \subset A \subset \lim_{m\to\infty} \text{INT}_n(A) \subset \overline{A}. \tag{4.6}$$

En primer lugar, vamos a probar la inclusión $\mathring{A} \subset \lim_{m\to\infty} \text{INT}_n(A)$. Para ello, consideramos $x \in \mathring{A}$. Por ser $\mathring{A}$ abierto, existe un radio r tal que la bola $B(x,r)$ está contenida en $\mathring{A}$. Por el Lema 4.1.1, se deduce que existen $n_0 \geq 1$ y $1 \leq k \leq m_{n_0}$ tales que $x \in T_{k,n_0} \subset B(x,r)$. En consecuencia, $T_{k,n_0} \subset \mathring{A} \subset A$ y por tanto $x \in \text{INT}_{n_0}(A) \subset \lim_{m\to\infty} \text{INT}_n(A)$.

Las inclusiones (4.3) permiten deducir

$$\lim_{m\to\infty} \text{INT}_n(A) = \bigcup_{n\geq 1} \text{INT}_n(A) \subset \bigcup_{n\geq 1} A = A,$$

$$A = \bigcap_{n\geq 1} A \subset \bigcap_{n\geq 1} \text{SUP}_n(A) = \lim_{m\to\infty} \text{SUP}_n(A).$$

Para demostrar la última inclusión de (4.6), vamos a probar la inclusión de los complementarios

$$\mathbb{R}^d \setminus \overline{A} \subset \mathbb{R}^d \setminus (\lim_{m\to\infty} \text{INT}_n(A)).$$

Sea $x \in \mathbb{R}^d \setminus \overline{A}$. Por ser $\mathbb{R}^d \setminus \overline{A}$ un conjunto abierto, existe un radio r tal que la bola $B(x,r)$ está contenida en $\mathbb{R}^d \setminus \overline{A}$. Por el Lema 4.1.1, se deduce que existen $n_0 \geq 1$ y $1 \leq k \leq m_{n_0}$ tales que $x \in T_{k,n_0} \subset B(x,r)$. En consecuencia,

$$T_{k,n_0} \subset \mathbb{R}^d \setminus \overline{A} \subset \mathbb{R}^d \setminus A$$

y por tanto $T_{k,n_0} \cap A = \emptyset$. Entonces, se deduce que $x \notin \text{SUP}_{n_0}(A)$ y por ello

$$x \notin \lim_{m\to\infty} \text{SUP}_n(A) \subset \text{SUP}_{n_0}(A).$$

Dado que la frontera de A tiene probabilidad nula, se deducen las igualdades

$$P(\mathring{A}) = P(A \setminus Fr(A)) = P(A) = P(A \cup Fr(A)) = P(\overline{A}),$$

y entonces se tiene que $P\left(\lim_{m\to\infty} \text{INT}_n(A)\right) = P(A) = P\left(\lim_{m\to\infty} \text{SUP}_n(A)\right)$ como consecuencia de las inclusiones (4.6). Entonces se tiene que

$$\lim_{m\to\infty} P_{\mathcal{T}_n}(\text{INT}_n(A)) = \lim_{m\to\infty} P(\text{INT}_n(A)) = P\left(\lim_{m\to\infty} \text{INT}_n(A)\right) = P(A)$$
$$\lim_{m\to\infty} P_{\mathcal{T}_n}(\text{SUP}_n(A)) = \lim_{m\to\infty} P(\text{SUP}_n(A)) = P\left(\lim_{m\to\infty} \text{SUP}_n(A)\right) = P(A)$$

de lo que se deduce la convergencia

$$|P_{\mathcal{T}_n}(A) - P(A)| \leq P_{\mathcal{T}_n}(\text{SUP}_n(A)) - P_{\mathcal{T}_n}(\text{INT}_n(A)) \xrightarrow[m\to\infty]{} 0.$$

En virtud del Teorema Portmanteau se concluye que la convergencia débil de $P_{\mathcal{T}_n}$ a P. □

Ejemplo: Un caso particular de discretización Euleriana surge de emplear una malla o partición regular formada por rectángulos d-dimensionales. Este ejemplo es muy significativo porque es fácil de implementar computacionalmente y aparece en situaciones prácticas como en el tratamiento de los colores de una imagen: cada píxel de la imagen está representado por una terna de números entre 0 y 255, las cuales representan la intensidad de los colores primarios R, G, B, rojo, verde y azul respectivamente. Si se considera la distribución de los colores en la imagen como una distribución en el cubo $[0, 255]^3$, entonces la información almacenada de la imagen no es más que una discretización Euleriana de la distribución real de los colores en una cuadrícula regular de 256^3 celdas. A continuación, se va a detallar formalmente la construcción de esta discretización en el caso general.

En primer lugar, fijamos un intervalo compacto I de $\mathbb{R}^d$ que contenga al soporte de P, es decir, un producto

$$I = I_1 \times \cdots \times I_d$$

de d intervalos cerrados y acotados de $\mathbb{R}$ a los cuales llamaremos lados de I. Escribiremos

$$I_i = [c_i, d_i]$$

para denotar los extremos de los intervalos I_i, $1 \leq i \leq d$. Para cada i, tomamos $m_i \geq 1$ y definimos los incrementos $h_i = \frac{d_i - c_i}{2^{m_i}}$. Consideramos la partición $\{I_i^{(j),m_i}\}_{j=1}^{2^{m_i}}$ del lado de I_i en 2^{m_i} subintervalos regulares, donde cada subintervalo $I_i^{(j),m_i}$ está definido por la siguiente expresión:

$$I_i^{(j),m_i} = \begin{cases} [c_i + (j-1)\cdot h_i,\ c_i + j \cdot h_i) & \text{si } 1 \leq j < 2^{m_i} \\ [c_i + (2^{m_i} - 1)\cdot h_i,\ d_i] & \text{si } j = 2^{m_i}. \end{cases}$$

Dado un vector de índices $\mathbf{j} = (j_1, \ldots, j_d) \in \{1, \ldots, 2^{m_1}\} \times \{1, \ldots, 2^{m_d}\} = J_{\mathbf{m}}$ escribimos

$$I^{\mathbf{j},\mathbf{m}} \overset{\text{def.}}{=} I_1^{(j_1),m_1} \times \cdots \times I_d^{(j_d),m_d}.$$

También emplearemos la notación $\mathbf{m} = (m_1, \ldots, m_d)$ para hacer referencia al número de subintervalos en los que se ha dividido cada lado de I. Lo que se ha conseguido es obtener una partición $\{I^{\mathbf{j},\mathbf{m}}\}_{\mathbf{j}\in J_{\mathbf{m}}}$ del intervalo I en $2^{m_1} \times \cdots \times 2^{m_d} = 2^{m_1+\cdots+m_d}$ subintervalos del mismo tamaño. Todos ellos tienen medida de Lebesgue

$$\lambda(I^{\mathbf{j},\mathbf{m}}) = h_1 \cdots h_d.$$

Si para cada $\mathbf{j} \in$ consideramos el punto

$$x^{\mathbf{j},\mathbf{m}} = (c_1 + (j_1 - 1)\cdot h_1, \ldots, c_m + (j_d - 1)\cdot h_d)$$

podemos observar que pertenece al intervalo $I^{\mathbf{j},\mathbf{m}}$ y en particular es una esquina de este. Con las notaciones anteriores podemos definir

$$\mathcal{T}_{\mathbf{m}} = \{(I^{\mathbf{j},\mathbf{m}}, x^{\mathbf{j},\mathbf{m}})\}_{\mathbf{j}\in J_{\mathbf{m}}}$$

la cual es una partición centrada del compacto I. Mediante el Teorema de Pitágoras, resulta sencillo calcular diámetro de cada intervalo $I^{\mathbf{j},\mathbf{m}}$ el cual vale $\delta(I^{\mathbf{j},\mathbf{m}}) = \sqrt{h_1^2 + \cdots + h_d^2}$. En consecuencia, se deduce que el diámetro de la partición es

$$\delta(\mathcal{T}_{\mathbf{m}}) = \sqrt{h_1^2 + \cdots + h_d^2}.$$

A partir de $\mathcal{T}_{\mathbf{m}}$ podemos construir la discretización Euleriana $P_{\mathcal{T}_{\mathbf{m}}}$ que hemos definido previamente. La denotaremos por $P_{\mathbf{m}}$ para aligerar la notación.

Si consideramos una sucesión de vectores de índices $\{\mathbf{m}_n\}_{n\geq 1}$ tal que la componente i-ésima de $\mathbf{m}_n$ es menor o igual que la componente i-ésima de $\mathbf{m}_{n+1}$ y además la sucesión $\min(\mathbf{m}_n)$ tiende a infinito, se verifica que $\{\mathcal{T}_{\mathbf{m}_n}\}_{n\geq 1}$ es una sucesión admisible de particiones centradas de I. Este hecho se puede comprobar notando que las particiones de cada lado de I que definen los subintervalos $I^{\mathbf{j},\mathbf{m}_{n+1}}$ son un refinamiento de las correspondientes particiones que definen los subintervalos $I^{\mathbf{j},\mathbf{m}_n}$. También se verifica que los centros de la partición $\mathcal{T}_{\mathbf{m}_n}$ son centros de $\mathcal{T}_{\mathbf{m}_n}$ por la definición de $x^{\mathbf{j},\mathbf{m}}$. Además, podemos acotar el diámetro de cada partición $\mathcal{T}_{\mathbf{m}}$:

$$\delta(\mathcal{T}_{\mathbf{m}}) = \sqrt{h_1^2 + \cdots + h_d^2} \leq \sqrt{d} \cdot \max_{1\leq i\leq d}(h_i) = \sqrt{d} \cdot \min_{1\leq i\leq d}\left(\frac{m_i}{c_i - d_i}\right) \leq \frac{\sqrt{d}}{\delta(I)} \cdot \min_{1\leq i\leq d}(m_i)\,.$$

Esta cota permite garantizar la convergencia del diámetro de las particiones $\mathcal{T}_{\mathbf{m}_n}$ a 0. Es decir, si refinamos la partición de cada lado de I según se ha descrito, estamos en condiciones de aplicar el Teorema 4.1.2. En esta situación, las discretizaciones Eulerianas $P_{\mathbf{m}_n}$ convergen débilmente a la probabilidad original P.

Observación 21. En el ejemplo de discretización Euleriana que se ha desarrollado, la probabilidad $P_{\mathbf{m}}$ se corresponde con discretizar la probabilidad original P en una cuadrícula uniforme. Si se quiere almacenar la información de $P_{\mathbf{m}}$, se debe almacenar un vector de dimensiones $2^{m_1} \times \cdots \times 2^{m_d}$ que se corresponde con las cantidades $P(I^{\mathbf{j},\mathbf{m}})$ para $\mathbf{j} \in J_{\mathbf{m}}$. Evidentemente, según se aumente el número de celdas de la cuadrícula, aumentará el gasto de almacenamiento de $P_{\mathbf{m}}$.

Debemos tener en cuenta que según se aumenta la dimensión del espacio en el que está definida la probabilidad P, el número de celdas necesarias en la discretización para obtener el mismo número de divisiones en cada dimensión crece exponencialmente. Esto supone un problema cuando se trabaja con conjuntos de datos que se representan con vectores de muchas dimensiones como puede suceder en problemas del área de *Machine Learning*. Además, en muchos casos prácticos, habrá muchas entradas de este vector que sean nulas o muy próximas a 0. Esta situación es ineficiente ya que la información obtenida no es muy representativa en comparación con el gasto computacional y de almacenamiento que supone emplear vectores tan masivos para trabajar con los datos.

4.1.2. Discretización Lagrangiana

La estrategia que se lleva a cabo en la discretización Lagrangiana es completamente distinta a la de la Euleriana. En este caso, suponemos que tenemos n realizaciones independientes $x_1, \ldots, x_m \in \mathbb{R}^d$ de un vector aleatorio con ley P. Consideramos la medida empírica P_m asociada a estas observaciones la cual viene dada por

$$P_m(A) = \frac{1}{m}\sum_{i=1}^{m} I_A(x_i). \tag{4.7}$$

Es decir, P_m es una probabilidad con soporte en los puntos x_i donde hemos considerado cada uno de ellos equiprobables. Evidentemente, la probabilidad resultante depende de las observaciones de P obtenidas. Podemos

hacer más explícito este componente aleatorio considerando m vectores aleatorios $X_1, \ldots, X_m$ independientes e igualmente distribuidos con ley P definidos en un mismo espacio Ω, y escribiendo P_m como

$$P_m^\omega(A) = \frac{1}{m} \sum_{i=1}^{m} I_A(X_i(\omega)) \quad \text{para cada } \omega \in \Omega. \tag{4.8}$$

Proposición 4.1.1. *Sea $\{X_n\}_{n\geq 1}$ una sucesión de vectores aleatorios independientes de $\mathbb{R}^d$ definidos en un mismo espacio probabilístico $(\Omega, \mathcal{F}, \nu)$. Supongamos que los vectores X_n son igualmente distribuidos con ley inducida P. Supongamos que el soporte de P está contenido en un compacto K. Fijado $\omega \in \Omega$, para cada $m \geq 1$ consideramos la discretización Lagrangiana P_m^ω dada por la expresión* (4.8)*. Entonces se tiene la convergencia débil*

$$P_m^\omega \xrightarrow{w} P \quad \nu \text{ - casi seguro.}$$

Demostración. El resultado que tenemos que probar es la existencia de un subconjunto Ω_0 de Ω tal que $\nu(\Omega_0) = 0$ y para cada $\omega \in \Omega \setminus \Omega_0$ se satsifaga $P_m^\omega \xrightarrow{w} P$.

Considerando una función f uniformemente continua y acotada en K, podemos ver que la integral de f respecto de P_m toma la siguiente expresión

$$\int_{\mathbb{R}^m} f(x)\, dP_m^\omega(x) = \frac{1}{m} \sum_{i=1}^{m} f(X_i(\omega)).$$

La Ley Fuerte de los Grandes Números nos asegura que $\frac{1}{m}\sum_{i=1}^m f(X_i)$ converge a $\int_{\mathbb{R}^d} f(x)\, dP(x)$ ν-casi seguro. Tomamos un subconjunto D denso y numerable de las funciones continuas y acotadas en K para la convergencia uniforme y para cada $f \in D$ consideramos Ω_f el conjunto de probabilidad nula para el cual no se da la convergencia de las integrales

$$\int_{\mathbb{R}^d} f(x)\, dP_m^\omega(x) \longrightarrow \int_{\mathbb{R}^d} f(x)\, dP(x). \tag{4.9}$$

El conjunto $\Omega_0 = \bigcup_{f\in D} \Omega_f$, es un conjunto de probabilidad nula y en su complementario, $\Omega \setminus \Omega_0$, se satisface (4.9) para cada $f \in D$.

Sea g una función continua y acotada en K. Entonces, por ser D denso, existe una sucesión $\{f_i\}_{i=1}^\infty$ en D que converge a g en la norma uniforme. Fijado $\varepsilon > 0$ consideramos un ínidice i para el cual $\|f_i - g\|_\infty < \frac{\varepsilon}{3}$. Como se ha probado, se satisface la igualdad (4.9) para f_i en el complementario de Ω_0. Por lo tanto, fijado $\omega \in \Omega \setminus \Omega_0$, existe $m_0 \geq 1$ tal que

$$\Big| \int_{\mathbb{R}^d} f_i(x)\, dP_m^\omega(x) - \int_{\mathbb{R}^d} f_i(x)\, dP(x) \Big| < \frac{\varepsilon}{3}$$

para cada $m \geq m_0$. De esta forma se tiene

$$\begin{aligned}
\Big| \textstyle\int g(x)\, dP_m^\omega(x) - \int g(x)\, dP(x) \Big| &\leq \Big| \textstyle\int g(x)\, dP_m^\omega(x) - \int f_i(x)\, dP_m^\omega(x) \Big| \\
+ \Big| \textstyle\int f_i(x)\, dP_m^\omega(x) - \int f_i(x)\, dP(x) \Big| &+ \Big| \textstyle\int f_i(x)\, dP(x) - \int g(x)\, dP(x) \Big| \\
\leq \textstyle\int |g(x) - f_i(x)|\, dP_m^\omega(x) + \Big| \int f_i(x)\, dP_m^\omega(x) - \int f_i(x)\, dP(x) \Big| &+ \int |f_i(x) - g(x)|\, dP(x) \\
\leq \|f_i - g\|_\infty \textstyle\int dP_m^\omega + \frac{\varepsilon}{3} + \leq \|f_i - g\|_\infty \int dP < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} < \varepsilon,
\end{aligned}$$

con lo que se prueba la convergencia (4.9) para g. Por el Teorema Portmanteau se deduce la convergencia débil de P_m^ω a P. □

Observación 22. Se ha probado la convergencia débil $P_m \xrightarrow{w} P$ de las probabilidades discretizadas a la probabilidad original en ambas configuraciones, Euleriana y Lagrangiana, bajo la hipótesis de que el soporte de P es compacto (en la configuración Lagrangina la convergencia débil se da casi seguro). En esta situación, para cada $p > 1$ se puede deducir la convergencia

$$\mathcal{W}_p(P_m, P) \xrightarrow[m\to\infty]{} 0$$

como consecuencia del Teorema 1.4.2. Fijado un punto x_0 del soporte de P, basta comprobar que $d(\cdot, x_0)^p$ es una función continua que está acotada en el soporte de P, por ser este un conjunto compacto. Entonces, la convergencia débil de las discretizaciones a P implica la convergencia

$$\int_{\mathbb{R}^n} d(x, x_0)^p \, dP_m = \int_{supp(P)} d(x, x_0)^p \, dP_m \xrightarrow[m\to\infty]{} \int_{supp(P)} d(x, x_0)^p \, dP = \int_{\mathbb{R}^n} d(x, x_0)^p \, dP.$$

Es decir, las discretizaciones se aproximan a la probabilidad original en la distancia p de Wasserstein.

4.2. Problemas de minimización en el espacio de Wasserstein

Se van a introducir dos problemas variacionales en el espacio de Wasserstein que pueden ser formulados como (4.1). Se estudiarán dos simplificaciones de cada uno de ellos correspondiendo con los dos enfoques que se han dado en la sección previa.

4.2.1. Problema de proyección

En primer lugar, vamos a introducir el problema de proyección de una probabilidad en un espacio $\mathcal{X}$ con momento de orden p finito en un subconjunto $\mathcal{S}$ de $\mathcal{P}_p(\mathcal{X})$. La motivación de este problema surge de querer aproximar una probabilidad por otra más sencilla que pertenezca a este subconjunto. Para que la probabilidad por la que aproximamos se ajuste al máximo a la probabilidad original empleamos la distancia de Wasserstein.

Vamos a estudiar el caso en el que el espacio $\mathcal{X}$ es $\mathbb{R}^k$, el cual va a ser el que aparezca en las situaciones prácticas. También, supondremos que el subconjunto está parametrizado por una aplicación $\alpha : \theta \mapsto \alpha(\theta)$ donde el parámetro θ pertenece a un abierto conexo $\Theta \subset \mathbb{R}^d$. Si consideramos una probabilidad β sobre $\mathbb{R}^k$ con momento de orden p finito, la formulación general del problema de proyección de β sobre el subconjunto parametrizado por α es

$$\arg\min_{\theta\in\Theta} \mathcal{W}_p(\alpha(\theta), \beta).$$

Podemos reconocer que este problema sigue la estructura (4.1) considerando el funcional $\mathcal{F}(\alpha) = \mathcal{W}_p(\alpha, \beta)$, el cual claramente involucra la distancia de Wasserstein. El conjunto $\{\alpha(\theta) : \theta \in \Theta\}$ de todas las probabilidades parametrizadas es el que juega el papel de $\mathcal{S}$ en la expresión (4.1). Notamos que si no imponemos ninguna condición en la parametrización α no tenemos garantías de la existencia del mínimo de $\mathcal{W}_p(\alpha(\theta), \beta)$.

Atendiendo a las razones que hemos expuesto en el Capítulo 3 y al inicio del Capítulo 3, vamos a estudiar una relajación del problema en la que sustituiremos la distancia p de Wasserstein por la distancia regularizada $\mathcal{W}_p^\varepsilon$. En estas condiciones el problema al que vamos a dedicar esta sección se formula de la siguiente manera:

$$arg\min_{\theta\in\Theta} \mathcal{W}_p^\varepsilon(\alpha(\theta), \beta), \tag{4.10}$$

o equivalentemente

$$\arg\min_{\theta\in\Theta} \mathcal{E}(\theta) \tag{4.11}$$

definiendo la función $\mathcal{E}(\theta) \overset{\text{def.}}{=} \mathcal{L}^{\varepsilon}_{\|\cdot\|_2^p}(\alpha(\theta), \beta)$. Podemos observar que ambas expresiones proporcionan la misma solución ya que si θ_0 minimiza (4.10), también minimiza la expresión (4.11) y viceversa.

Vamos a estudiar dos simplificaciones del problema de proyección las cuales están basadas en los dos tipos de discretizaciones que hemos estudiado. Estos dos enfoques permiten emplear herramientas del cálculo diferencial para obtener la solución del problema. En ambos casos, vamos a considerar que la probabilidad β tiene soporte finito formado por n puntos $\{y_1, \ldots, y_n\}$. Para ello, debemos discretizar β con cualquiera de los esquemas que hemos introducido. De esta forma, podemos aproximar la probabilidad β por una probabilidad con soporte finito de la siguiente manera

$$\beta = \sum_{j=1}^{n} b_j \delta_{y_j}, \tag{4.12}$$

y emplearemos el vector $\mathbf{b} = (b_1, \ldots, b_n)$ para representar a β.

En las dos situaciones que vamos a estudiar, la configuración Euleriana y la configuración Lagrangiana, vamos a hacer uso del Algoritmo de Descenso de gradiente (ver Apéndice B) para encontrar una solución del problema de proyección. Para poder emplear esta herramienta es necesario que la función $\mathcal{E}$ sea continuamente diferenciable. Esta razón motiva que estudiemos la regularidad que debemos exigir a las parametrizaciones para garantizar que $\mathcal{E}$ sea de clase $\mathcal{C}^1$ en cada una de las configuraciones.

Configuración Euleriana: En primer lugar, vamos a estudiar la configuración Euleriana del problema de proyección. Esta consiste en considerar una malla finita de celdas y un centro en cada celda de forma que todas las probabilidades $\alpha(\theta)$ tengan soporte contenido en el conjunto de centros. La manera de formular el problema formalmente consiste en fijar un compacto K y una partición centrada de K

$$\mathcal{T} = \{(T_k, x_k)\}_{k=1}^{m}$$

formada por m celdas y m centros. Vamos a consideramos que las probabilidades $\alpha(\theta)$ tienen soporte discreto contenido en $X = \{x_1, \ldots, x_m\}$. Con esta simplificación, podemos obtener una expresión sencilla para la distancia de Wasserstein: para cada $\theta \in \Theta$, consideramos el vector $\mathbf{a}(\theta)$ de orden m cuya componente i-ésimas es la probabilidad puntual $a_i(\theta) = \alpha(\theta)(x_i)$ para $1 \leq i \leq m$. Es decir, podemos escribir la probabilidad $\alpha(\theta)$ como

$$\alpha(\theta) = \sum_{i=1}^{m} a_i(\theta) \delta_{x_i}.$$

Denotamos por $\mathbf{C}$ a la matriz de orden $m \times n$ cuyas componentes vienen dadas por

$$c_{i,j} = \|x_i - y_j\|_2^p,$$

es decir, $c_{i,j}$ es la distancia entre los puntos x_i e y_j elevada a p. Con esta observación podemos reescribir el problema de proyección (4.11) adaptándolo a la configuración Euleriana de la siguiente manera:

$$\arg\min_{\theta \in \Theta} \mathcal{E}_E(\theta), \tag{4.13}$$

donde definimos $\mathcal{E}_E(\theta) \overset{\text{def.}}{=} \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}(\theta), \mathbf{b})$. El problema de minimización anterior se puede escribir como (4.2) considerando el funcional regularizado $\tilde{\mathcal{F}}(\alpha) = \left(\mathcal{W}_2^{\varepsilon}(\alpha, \beta)\right)^2$ donde β está dada por (4.12). El conjunto $\mathcal{S}_F$ que interviene en la formulación general (4.2) viene dado por

$$\{\textstyle\sum_{i=1}^{m} a_i(\theta) \delta_{x_i} : \theta \in \Theta\},$$

es decir, el conjunto de probabilidades con soporte contenido en $\{x_1, \ldots, x_m\}$ parametrizadas por la aplicación $\mathbf{a} : \theta \mapsto \mathbf{a}(\theta)$.

Observación 23. Podemos observar que la función $\mathcal{E}_L$ es la composición de tres aplicaciones:

- $\mathbf{a} : \theta \mapsto \mathbf{a}(\theta)$, que parametriza el vector de probabilidades puntuales $\mathbf{a}(\theta) \in \Sigma_m$.
- $\mathbf{a} \mapsto (\mathbf{a}, \mathbf{b})$, que forma un vector de $\Sigma_m \times \Sigma_n$ a partir de un vector $\mathbf{a} \in \Sigma_m$ añadiendo las componentes del vector $\mathbf{b}$.
- $(\mathbf{a}, \mathbf{b}) \mapsto \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$, la cual obtiene el coste de transporte entrópico discreto entre las probabilidades dadas por $\mathbf{a}$ y $\mathbf{b}$ con matriz de coste $\mathbf{C}$ y factor de regularización ε.

Claramente la segunda de estas aplicaciones es diferenciable en la variedad Σ_m y en virtud de la Proposición 3.3.1 podemos asegurar que la tercera aplicación es continuamente diferenciable como función definida en la variedad $\Sigma_m \times \Sigma_n$. Supongamos que la parametrización $\mathbf{a} : \theta \mapsto \mathbf{a}(\theta) \in \Sigma_m$ es de clase $\mathcal{C}^1$ en Θ. Entonces, por la regla de la cadena deducimos que la función $\mathcal{E}_E$ es continuamente diferenciable en Θ y su gradiente es

$$\nabla \mathcal{E}_E(\theta) = \nabla_{\mathbf{a},\mathbf{b}} \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}(\theta), \mathbf{b}) \cdot \begin{bmatrix} Id_m \\ 0_{m\times n} \end{bmatrix} \cdot \frac{d\mathbf{a}}{d\theta}(\theta) = (\mathbf{f}(\theta), \mathbf{g}(\theta)) \cdot \begin{bmatrix} Id_m \\ 0_{m\times n} \end{bmatrix} \cdot \frac{d\mathbf{a}}{d\theta}(\theta) = \mathbf{f}(\theta) \cdot \frac{d\mathbf{a}}{d\theta}(\theta), \quad (4.14)$$

donde $(\mathbf{f}(\theta), \mathbf{g}(\theta))$ son los potenciales en los que se alcanza el óptimo del problema dual de transporte entrópico discreto entre las probabilidades $\alpha(\theta)$ y β, las cuales vienen determinadas por los vectores $\mathbf{a}(\theta)$ y $\mathbf{b}$.

La expresión (4.14) del gradiente de $\mathcal{E}_E$ permite implementar el Algoritmo de Descenso de gradiente para obtener una solución del problema de proyección en la configuración Euleriana.

Observación 24. Un caso particular de esta situación que merece un estudio particular es cuando se considera una parametrización $\mathbf{a}$ lineal y Θ es un subconjunto convexo de $\mathbb{R}^d$. En esta situación se tiene que

$$\mathbf{a}(\lambda\theta + (1-\lambda)\theta') = \lambda\mathbf{a}(\theta) + (1-\lambda)\mathbf{a}(\theta')$$

para cada $\lambda \in [0,1]$ y por la convexidad de $\mathcal{L}^{\varepsilon}_c$ que se ha probado en la Observación 13, se deduce la convexidad de $\mathcal{E}_E$:

$$\begin{aligned}\mathcal{E}_E(\lambda\theta + (1-\lambda)\theta') &= \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}(\lambda\theta + (1-\lambda)\theta'), \mathbf{b}) = \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\lambda\mathbf{a}(\theta) + (1-\lambda)\mathbf{a}(\theta'), \mathbf{b}) \\ &\leq \lambda\mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}(\theta), \mathbf{b}) + (1-\lambda)\mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}(\theta'), \mathbf{b}) = \lambda\mathcal{E}_E(\theta) + (1-\lambda)\mathcal{E}_E(\theta').\end{aligned}$$

Esta propiedad nos asegura que todo mínimo local de $\mathcal{E}_E$ en Θ es extremo absoluto. Entonces, si el Algoritmo de Descenso de gradiente converge, la solución obtenida será un mínimo de $\mathcal{E}_E$. Este resultado se cumple en todos los problemas de optimización convexa y se demuestra en el siguiente lema.

Lema 4.2.1. Sea C un conjunto convexo de $\mathbb{R}^d$ y sea $f : C \to \mathbb{R}$ una función convexa. Entonces, todo extremo local de f en C es extremo absoluto de f.

Demostración. Se va probar el resultado por reducción al absurdo, por lo que vamos a suponer que la propiedad enunciada es falsa. Entonces existen dos puntos $x_0, x_0' \in C$ y una bola centrada en x_0, $B(x_0, r)$, tales que $f(x) \geq f(x_0)$ para cada $x \in B(x_0, r) \cap C$ y además $f(x_0') < f(x_0)$. Por convexidad de C, se tiene que el segmento que une x_0 con x_0' está contenido en C. Por lo tanto, existe un $\lambda \in (0,1)$ para el cual $\lambda x + (1-\lambda)x_0' \in B(x_0, r) \cap C$ y por la convexidad de f se deduce que

$$f(\lambda x_0 + (1-\lambda)x_0') \leq \lambda f(x_0) + (1-\lambda)f(x_0') < \lambda f(x_0) + (1-\lambda)f(x_0) = f(x_0),$$

lo que supone una contradicción. □

Un ejemplo de la situación descrita en la Observación 24 es el siguiente: tomando $\Theta = \Sigma_m$ y la parametrización $\mathbf{a} : \theta \mapsto \mathbf{a}(\theta) = \theta$ dada por la identidad, entonces se tiene $\mathcal{E}_E(\mathbf{a}) = \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$ para $\mathbf{a} \in \Sigma_m$ y por lo tanto el problema de proyección se escribe

$$\arg\min_{\mathbf{a}\in\Sigma_m} \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}).$$

El gradiente de la función objetivo es $\nabla \mathcal{E}_E(\mathbf{a}) = \mathbf{f}(\mathbf{a})$ el potencial óptimo del problema dual de transporte óptimo.

Configuración Lagrangiana: La otra simplificación del problema de proyección que vamos a estudiar emplea un enfoque muy distinto al Euleriano. En la configuración Lagrangiana en vez de fijar los puntos del soporte de las probabilidades $\alpha(\theta)$ se considera que estas son probabilidades empíricas, es decir, que $\alpha(\theta)$ es uniforme en su soporte, el cual supondremos que está formado por m puntos distintos. De esta forma, la probabilidad $\alpha(\theta)$ que se puede expresar como

$$\alpha(\theta) = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i(\theta)}.$$

Para cada $\theta \in \Theta$ consideramos la matriz de coste $\mathbf{C}(\theta)$ de orden $m \times n$ cuyas componentes vienen dadas por

$$c_{i,j}(\theta) = \|x_i(\theta) - y_j\|_2^p.$$

De esta forma, considerando la función $\mathcal{E}_L(\theta) \overset{\text{def.}}{=} \mathcal{L}^{\varepsilon}_{\mathbf{C}(\theta)}(\mathbb{1}/m, \mathbf{b})$, la formulación del problema de proyección en la configuración Lagrangiana es

$$\arg\min_{\theta \in \Theta} \mathcal{E}_L(\theta). \tag{4.15}$$

La expresión previa se corresponde con un problema de minimización como el descrito en (4.2). Para comprobarlo basta considerar el conjunto

$$\mathcal{S}_F = \{\tfrac{1}{m} \textstyle\sum_{i=1}^{m} \delta_{x_i(\theta)}\}$$

dado por las probabilidades empíricas con soporte parametrizado por la aplicación $\mathbf{x} : \theta \mapsto \mathbf{x}(\theta)$, donde $\mathbf{x}(\theta) = (x_1(\theta), \ldots, x_m(\theta))$. El funcional regularizado es $\tilde{\mathcal{F}}(\alpha) = (\mathcal{W}_2^{\varepsilon}(\alpha, \beta))^2$ donde β está dada por (4.12).

Observación 25. Vamos a estudiar las condiciones que debemos imponer en la parametrización para que la función $\mathcal{E}_L$ sea regular. Para ello, consideramos las siguientes aplicaciones:

- $\mathbf{x} : \theta \mapsto \mathbf{x}(\theta) = (x_1(\theta), \ldots, x_m(\theta))$ que asigna a cada parámetro $\theta \in \Theta$ el soporte de la probabilidad $\alpha(\theta)$.
- $\mathbf{C} : \mathbf{x} \mapsto \mathbf{C}(\mathbf{x})$ donde $\mathbf{C}$ es la matriz de coste de orden $m \times n$ que viene dada por

$$c_{i,j}(\mathbf{x}) = \|x_i - y_j\|_2^p.$$

- $\mathcal{R} : \mathbf{C} \mapsto \mathcal{R}(\mathbf{C}) = \mathcal{L}^{\varepsilon}_{\mathbf{C}}(\mathbb{1}/m, \mathbf{b})$ que obtiene el coste de transporte entrópico discreto a partir de la matriz de coste de orden $m \times n$.

Podemos notar que $\mathcal{E}_L$ es la composición de estas tres aplicaciones

$$\mathcal{E}_L(\theta) = \mathcal{R}(\mathbf{C}(\mathbf{x}(\theta))).$$

Es fácil ver que la aplicación $\mathbf{C}$ es continuamente diferenciable en $\mathbb{R}^m$ y por la Proposición 3.3.2 también podemos asegurar que $\mathcal{R}$ es de clase $\mathcal{C}^1$ en $\mathbb{R}^{m \times n}$. Si suponemos que la aplicación $\mathbf{x}$ es continuamente diferenciable en Θ, entonces por la regla de la cadena deducimos que $\mathcal{E}_L$ también es continuamente diferenciable en Θ. Además, en esta situación podemos obtener una expresión para el gradiente de $\mathcal{E}_L$:

$$\nabla \mathcal{E}_L(\theta) = \nabla \mathcal{R}(\mathbf{C}(\mathbf{x}(\theta))) \cdot \frac{d\mathbf{C}}{d\mathbf{x}}(\mathbf{x}(\theta)) \cdot \frac{d\mathbf{x}}{d\theta}(\theta) = \mathbf{P}^{\varepsilon}(\mathbf{C}(\mathbf{x}(\theta))) \cdot \frac{d\mathbf{C}}{d\mathbf{x}}(\mathbf{x}(\theta)) \cdot \frac{d\mathbf{x}}{d\theta}(\theta). \tag{4.16}$$

Escribiremos $\mathbf{P}^{\varepsilon}(\theta)$ para denotar a la matriz $\mathbf{P}^{\varepsilon}(\mathbf{C}(\mathbf{x}(\theta)))$ y simplificar la notación. Vamos a dar la expresión de la matriz jacobiana

$$\frac{d\mathbf{C}}{d\mathbf{x}}(\mathbf{x}) = \left(\frac{\partial c_{i,j}}{\partial x_{s,t}}(\mathbf{x}) \right)_{\substack{(i,j) \in \{1,\ldots,m\} \times \{1\ldots,n\} \\ (s,t) \in \{1,\ldots,m\} \times \{1\ldots,k\}}},$$

la cual es una matriz de $\mathbb{R}^{mn\times mk}$, a partir de la definición de la aplicación $\mathbf{C}$. Dado un punto $x_s \in \mathbb{R}^k$, denotaremos por $x_{s,t}$ a su componente t-ésima para $1 \leq t \leq k$. Entonces, derivando la componente $c_{i,j}$ respecto de $x_{s,t}$ obtenemos

$$\begin{aligned}\frac{\partial c_{i,j}}{\partial x_{s,t}}(\mathbf{x}) &= \frac{\partial \|x_i - y_j\|_2^p}{\partial x_{s,t}}(\mathbf{x}) = p \cdot \|x_i - y_j\|_2^{p-1} \cdot \frac{1}{2\sqrt{\sum_{r=1}^k (x_{i,r} - y_{j,r})^2}} \cdot 2\sum_{r=1}^k (x_{i,r} - y_{j,r}) \cdot \frac{\partial x_{i,r}}{\partial x_{s,t}}(\mathbf{x}) \\ &= \frac{\partial \|x_i - y_j\|_2^p}{\partial x_{s,t}}(\mathbf{x}) = p \cdot \|x_i - y_j\|_2^{p-2} \cdot \sum_{r=1}^k (x_{i,r} - y_{j,r}) \cdot \delta_{(s,t)}(i,r)\end{aligned}$$

En consecuencia, la derivada parcial vale

$$\frac{\partial c_{i,j}}{\partial x_{s,t}}(\mathbf{x}) = \begin{cases} 0 & \text{si } i \neq s, \\ p \cdot \|x_i - y_j\|_2^{p-2} \cdot (x_{s,t} - y_{j,t}) & \text{si } i = s. \end{cases} \tag{4.17}$$

Teniendo en cuenta las expresiones(4.16) y (4.17) que hemos calculado, obtenemos una fórmula para el gradiente de la función $\mathcal{E}_L$.

$$\begin{aligned}\nabla \mathcal{E}_L(\theta) &= \left(\sum_{i,j} p_{i,j}^{\varepsilon}(\theta) \cdot \sum_t p \cdot \|x_i(\theta) - y_j\|_2^{p-2} \cdot (x_{i,t}(\theta) - y_{j,t}) \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) \right)_{1\leq l\leq d} \\ &= p \left(\sum_{i,j} p_{i,j}^{\varepsilon}(\theta) \cdot \|x_i(\theta) - y_j\|_2^{p-2} \cdot \sum_t \cdot (x_{i,t}(\theta) - y_{j,t}) \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) \right)_{1\leq l\leq d}\end{aligned} \tag{4.18}$$

En el caso particular $p = 2$, el gradiente de $\mathcal{E}_L$ adopta una expresión más sencilla:

$$\begin{aligned}\nabla \mathcal{E}_L(\theta) &= 2 \left(\sum_{i,j} p_{i,j}^{\varepsilon}(\theta) \cdot \sum_t \cdot (x_{i,t}(\theta) - y_{j,t}) \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) \right)_{1\leq l\leq d} \\ &= 2 \left(\sum_{i,t} x_{i,t}(\theta) \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) \cdot \sum_j p_{i,j}^{\varepsilon}(\theta) - \sum_{i,j,t} p_{i,j}^{\varepsilon}(\theta) \cdot y_{j,t} \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) \right)_{1\leq l\leq d} \\ &= 2 \left(\frac{1}{m} \sum_{i,t} x_{i,t}(\theta) \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) - \sum_{i,j,t} p_{i,j}^{\varepsilon}(\theta) \cdot y_{j,t} \cdot \frac{dx_{i,t}}{d\theta_l}(\theta) \right)_{1\leq l\leq d}\end{aligned}$$

La expresión (4.18) que hemos obtenido para el gradiente de la función $\mathcal{E}_L$ cuando la parametrización $\mathbf{x}$ es continuamente diferenciable, nos permite aplicar el Algoritmo de Descenso de gradiente (Algoritmo 4) para obtener la solución al problema de proyección en la configuración Lagrangiana. Sin embargo, la función $\mathcal{E}_L$ no es convexa por lo general, al contrario de lo que ocurría en la configuración Euleriana con parametrización lineal. Esto nos indica que pueden existir mínimos locales de $\mathcal{E}_L$ que no sean mínimos absolutos y por lo tanto el Algoritmo de Descenso de gradiente puede converger hacia uno de estos mínimos locales o hacia un punto de silla de $\mathcal{E}_L$.

Observación 26. El cálculo de los potenciales $(\mathbf{f}(\theta), \mathbf{g})$ y de la matriz $\mathbf{P}^{\varepsilon}(\theta)$, que intervienen en los gradientes de las funciones $\mathcal{E}_E$ y $\mathcal{E}_L$ respectivamente, requiere de obtener la solución de un problema de transporte entrópico discreto empleando el Algoritmo de Sinkhorn. Este cálculo se debe repetir en cada iteración del Descenso de gradiente para obtener una solución del problema de proyección en ambas configuraciones.

La observación previa es la que motiva el empleo de la Diferenciación Automática para el cálculo de los gradientes de las funciones $\mathcal{E}_E$ y $\mathcal{E}_L$ evitando calcular los vectores $(\mathbf{f}(\theta), \mathbf{g})$ o la matriz $\mathbf{P}^{\varepsilon}(\theta)$. En el Apéndice

C se introduce esta técnica que se emplea para el cálculo de matrices Jacobianas de aplicaciones que son composición de aplicaciones elementales. Para poder aplicar esta técnica al problema de proyección Euleriano se debe aproximar

$$\mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}(\theta), \mathbf{b}) \approx \langle \mathbf{P}^{(l)}, \mathbf{C} \rangle + \varepsilon \langle \mathbf{P}^{(l)}, \log(\mathbf{P}^{(l)}) - \mathbb{1} \rangle$$

donde $\mathbf{P}^{(l)}$ es la matriz resultante de la iteración l-ésima del Algoritmo de Sinkhorn para el cálculo de $\mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}(\theta), \mathbf{b})$. Con esta simplificación, la función $(\mathbf{a}, \mathbf{b}) \mapsto \mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b})$ es composición de funciones elementales. De forma similar, para poder utilizar la Diferenciación Automática en el cálculo del gradiente $\nabla \mathcal{E}_L$, debemos aproximar

$$\mathcal{L}_{\mathbf{C}}(\mathbf{x})^{\varepsilon}(\mathbb{1}/m, \mathbf{b}) \approx \langle \mathbf{P}^{(l)}, \mathbf{C}(\mathbf{x}) \rangle + \varepsilon \langle \mathbf{P}^{(l)}, \log(\mathbf{P}^{(l)}) - \mathbb{1} \rangle$$

donde $\mathbf{P}^{(l)}$ es la matriz resultante de la iteración l-ésima del Algoritmo de Sinkhorn para el cálculo de $\mathcal{L}_{\mathbf{C}}(\mathbf{x})^{\varepsilon}(\mathbb{1}/m, \mathbf{b})$. De esta manera, la función $\mathcal{R} : \mathbf{C} \mapsto \mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbb{1}/m, \mathbf{b})$ es composición de funciones elementales. El uso de la Diferenciación Automática, el cual está implementado en numerosas librerías, permite reducir el tiempo de computo de los gradientes y la aceleración del Algoritmo de Descenso de gradiente.

4.2.2. Baricentros en el espacio de Wasserstein

El segundo problema variacional en el espacio de Wasserstein que vamos a estudiar en es el del cálculo de baricentros de Wasserstein. Un baricentro de Wasserstein de s probabilidades $\beta_1, \ldots, \beta_s$ sobre un espacio $\mathcal{X}$ es una probabilidad que se aproxima lo máximo posible a $\beta_1, \ldots, \beta_s$ simultáneamente. Para cuantificar cuánto se debe asemejar el baricentro a cada una de estas probabilidades β_i empleamos pesos positivos λ_i.

Definición 4.2.1. Sean $\beta_1, \ldots \beta_s$ probabilidades sobre $\mathcal{X}$ con momento de orden p finito y sean $\lambda_1, \ldots, \lambda_s$ escalares positivos tales que $\sum_{t=1}^{s} \lambda_t = 1$. Decimos que una probabilidad $\alpha \in \mathcal{P}_p(\mathcal{X})$ es un baricentro de $\beta_1, \ldots \beta_s$ con pesos $\lambda_1, \ldots, \lambda_s$ si α es una solución de

$$\arg \min_{\alpha \in \mathcal{P}_p(\mathcal{X})} \sum_{t=1}^{s} \lambda_t \mathcal{W}_p^2(\alpha, \beta_t). \tag{4.19}$$

Resulta sencillo comprobar que el problema de 1 baricentro en el Espacio de Wasserstein es de la forma (4.1): tomando $\mathcal{F}(\alpha) = \sum_{t=1}^{s} \lambda_t \mathcal{W}_p^2(\alpha, \beta_t)$ y $\mathcal{S} = \mathcal{P}_p(\mathcal{X})$ se comprueba que verifica las condiciones exigidas.

Observación 27. Podemos entender un baricentro en el espacio de Wasserstein como una media de Fréchet. Estas se definen sobre un espacio métrico $(\mathcal{Y}, d)$ de la siguiente manera: dados s puntos $y_1 \ldots, y_s$ de $\mathcal{Y}$ y $\lambda_1, \ldots, \lambda_s$ escalares positivos tales que $\sum_{t=1}^{s} \lambda_t = 1$, una solución de

$$\arg \min_{x \in \mathcal{Y}} \sum_{t=1}^{s} \lambda_t d(x, y_i)^2$$

es una media de Fréchet de $y_1 \ldots, y_s$ con pesos $\lambda_1, \ldots, \lambda_s$. En el caso que vamos a estudiar, la distancia escogida es la distancia p de Wasserstein.

Vamos a centrarnos en el caso en que $\mathcal{X} = \mathbb{R}^k$ y $p = 2$ de forma que estudiaremos el problema del cálculo de baricentro

$$\arg \min_{\alpha \in \mathcal{P}_2(\mathbb{R}^k)} \sum_{t=1}^{s} \lambda_t \mathcal{W}_2^2(\alpha, \beta_t)$$

dados $\beta_1, \ldots, \beta_s \in \mathcal{P}_2(\mathbb{R}^k)$ y pesos $(\lambda_1, \ldots, \lambda_s) \in \Sigma_s$. Esta decisión la tomamos en virtud del siguiente resultado, el cual nos garantiza la existencia de baricentros bajo estas condiciones.

Teorema 4.2.2. *Sean $\beta_1, \ldots, \beta_s \in \mathcal{P}_2(\mathbb{R}^k)$ probabilidades de $\mathbb{R}^k$ con momento de orden 2 finito y sean $\lambda_1, \ldots, \lambda_s$ escalares positivos tales que $\sum_{t=1}^{s} \lambda_t = 1$. Entonces existe un baricentro de $\beta_1, \ldots, \beta_s$ con pesos $\lambda_1, \ldots, \lambda_s$.*

Demostración. Este resultado se corresponde con la Proposición 2.3 de [14] donde se puede encontrar la demostración. □

En el Teorema 4.2.2 hablamos de la existencia de un baricentro ya que en principio no tenemos garantías de que este sea único. En [14] se estudia una condición sobre las probabilidades $\beta_1 \dots, \beta_s$ que asegura la unicidad del baricentro en el espacio $(\mathcal{P}_2(\mathbb{R}^k), \mathcal{W}_2)$.

Teorema 4.2.3. *Sean $\beta_1, \dots, \beta_s \in \mathcal{P}_2(\mathbb{R}^k)$ probabilidades de $\mathbb{R}^k$ con momento de orden 2 finito tales que existe al menos un índice $1 \leq t \leq s$ para el cual β_t no da probabilidad a subespacios de dimensión menor o igual a $n-1$. Sean $\lambda_1, \dots, \lambda_s$ escalares positivos tales que $\sum_{t=1}^{s} \lambda_t = 1$. Entonces existe un único baricentro de $\beta_1, \dots, \beta_s$ con pesos $\lambda_1, \dots, \lambda_s$.*

Demostración. La demostración del resultado se puede encontrar en [14, Prop.3.5]. □

El cálculo exacto de baricentros en el espacio de Wasserstein es complejo y por lo general no se tiene una expresión exacta. Su cálculo se realiza mediante procesos iterativos. En determinados casos particulares, se conocen expresiones para el baricentro de $\beta_1, \dots, \beta_s$. Uno de ellos se tiene cuando se consideran probabilidades continuas $\beta_1, \dots, \beta_s$ sobre $\mathbb{R}$ con momento de orden 2 finito. En este caso, el baricentro es único y se puede expresar como una combinación lineal de leyes inducidas. La demostración del resultado, la cual se puede encontrar en [14, Sec.6.1], se basa en la existencia de las aplicaciones $T_{1,i}$ de transporte óptimo de Monge entre β_1 y cada β_i. Como se ha visto en la Observación 6, se tiene la expresión exacta de estas aplicaciones lo que permite dar la fórmula del baricentro:

$$\left(\sum_{t=1}^{s} \lambda_t T_{1,i}\right) \#\, \beta_1 \quad \text{con } T_{1,i} = F_i^{-1} \circ F_1,$$

donde F_i y F_i^{-1} son la función de distribución y la función cuantil de β_i respectivamente.

Otro caso particular en el que se conoce la distribución del baricentro es cuando las probabilidades $\beta_1, \dots, \beta_s$ son distribuciones gaussianas. En esta situación, el baricentro de estas probabilidades, el cual es único, también tiene distribución gaussiana como se prueba en [14, Thm.6.1]. En [3] se desarrolla el cálculo de este baricentro a partir de los vectores de medias y las matrices de covarianzas de $\beta_1, \dots, \beta_s$ mediante un proceso iterativo basado en operaciones matriciales.

El enfoque que vamos a adoptar en esta sección se basa en estudiar la versión regularizada del problema de cálculo de baricentros en el espacio de Wasserstein. Esta decisión se toma para aprovechar la ventaja que aporta la convexidad estricta del transporte entrópico. Además, atendiendo a las observaciones dadas al inicio de capítulo, se centrará el estudio de baricentros regularizados al caso discreto con el fin de poder emplear técnicas del cálculo matricial.

Definición 4.2.2. Sean $\beta_1, \dots \beta_s$ probabilidades sobre $\mathcal{X}$ con momento de orden 2 finito, sean $(\lambda_1, \dots, \lambda_s) \in \Sigma_s$ pesos y $\varepsilon > 0$ un factor de regularización. Dada una probabilidad $\alpha \in \mathcal{P}_2(\mathcal{X})$ que es una solución de

$$\arg \min_{\alpha \in \mathcal{P}_2(\mathcal{X})} \sum_{t=1}^{s} \lambda_t (\mathcal{W}_2^{\varepsilon}(\alpha, \beta_t))^2 \tag{4.20}$$

decimos que α es un baricentro regularizado de $\beta_1, \dots \beta_s$ con pesos $\lambda_1, \dots, \lambda_s$.

A continuación introducimos la formulación del problema de baricentros regularizados en el caso discreto empleando la configuración Euleriana. Vamos a dedicar el resto de la sección al estudio en profundidad de este problema.

Configuración Euleriana: La simplificación que vamos a adoptar es similar a la que hemos hecho en el problema de proyección. Vamos a considerar que las probabilidades β_t tienen soporte finito $\{y_{t,1}, \dots, y_{t,n_t}\}$ formado por n_t puntos de $\mathbb{R}^k$ y las representaremos por vectores de probabilidades $\mathbf{b}_t \in \Sigma_{n_t}$. Además, vamos a fijar m puntos $\{x_1, \dots, x_m\}$ que suponemos que contienen el soporte de las probabilidades α entre las cuales vamos a buscar el baricentro. Es decir, vamos a representar cada probabilidad α por un vector $\mathbf{a} \in \Sigma_m$ y para cada t vamos a considerar las matrices de coste $\mathbf{C}_t$ cuyas componentes son

$$c_{t,i,j} = \|x_i - y_{t,j}\|_2^2.$$

Con estas notaciones el problema del baricentro regularizado en la configuración Euleriana es

$$\arg \min_{\mathbf{a} \in \Sigma_m} \sum_{t=1}^s \lambda_t \mathcal{L}^\varepsilon_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t).$$

Diremos que una solución $\mathbf{a}$ de este problema es un baricentro regularizado de $\mathbf{b}_1, \dots, \mathbf{b}_s$ para los pesos $\lambda_1, \dots, \lambda_s$. Debemos tener en consideración que estamos asumiendo que el baricentro regularizado de las probabilidades $\beta_1, \dots, \beta_s$ tiene soporte contenido en $\{x_1, \dots, x_m\}$. Esta suposición no se satisface por lo general, de forma que lo que se está calculando realmente es una proyección del baricentro regularizado sobre el conjunto las probabilidades parametrizadas por $\mathbf{a} \mapsto \sum_i^m a_i \delta_{x_i}$, $\mathbf{a} \in \Sigma_m$.

Podemos obtener una problema equivalente al baricentro regularizado que involucre las matrices en las que se alcanzan los óptimos de los costes entrópicos $\mathcal{L}^\varepsilon_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t)$ para cada $\mathbf{a} \in \Sigma_m$. Para ello, definimos primero el conjunto de matrices que pueden ser factibles para este problema.

Definición 4.2.3. Sean $\mathbf{b}_1, \dots, \mathbf{b}_s$ vectores de probabilidades tales que $\mathbf{b}_t \in \mathbb{R}^{n_t}$ para cada t y sea $m \geq 1$. Definimos el conjunto de matrices

$$\mathcal{U}(*, (\mathbf{b}_t)_t) = \{(\mathbf{P}_t)_t : \mathbf{P}_t \in \mathbb{R}^{m \times n_t},\ \mathbf{P}_1 \mathbb{1} = \cdots = \mathbf{P}_s \mathbb{1},\ (\mathbf{P}_t)^T \mathbb{1} = \mathbf{b}_t,\ 1 \leq t \leq s\}.$$

Lema 4.2.4. Sean $\mathbf{b}_1, \dots, \mathbf{b}_s$ vectores de probabilidades tales que $\mathbf{b}_t \in \mathbb{R}^{n_t}$ para cada t. Sea $\mathbf{a} \in \Sigma_m$ un baricentro regularizado de estas probabilidades con pesos $(\lambda_1, \dots, \lambda_s) \in \Sigma_s$ y factor de regularización $\varepsilon > 0$. Para cada t, $1 \leq t \leq s$, consideramos la matriz $\mathbf{P}_t^\varepsilon \in \mathcal{U}(\mathbf{a}, \mathbf{b}_t)$ en la que se alcanza $\mathcal{L}^\varepsilon_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t)$. Se tiene que en $(\mathbf{P}_t^\varepsilon)_t$ se alcanza el mínimo del conjunto

$$\Big\{ \sum_{t=1}^s \lambda_t \left(\langle \mathbf{P}_t, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t, \log(\mathbf{P}_t) - 1 \rangle \right) : (\mathbf{P}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t) \Big\}. \tag{4.21}$$

Demostración. Para demostrar este resultado, en primer lugar notamos que $\mathbf{P}_1^\varepsilon \mathbb{1} = \cdots = \mathbf{P}_s^\varepsilon \mathbb{1} = \mathbf{a}$ y $(\mathbf{P}_t^\varepsilon)^T \mathbb{1} = \mathbf{b}_t$ para cada t, $1 \leq t \leq s$. Esto se debe a que $\mathbf{P}_t^\varepsilon \in \mathcal{U}(\mathbf{a}, \mathbf{b}_t)$ para cada t.

Razonando por reducción al absurdo, podemos asumir que existen matrices $(\tilde{\mathbf{P}}_t)_t$ que pertenecen al conjunto $\mathcal{U}(*, (\mathbf{b}_t)_t)$ y además satisfacen

$$\sum_{t=1}^s \lambda_t \left(\langle \tilde{\mathbf{P}}_t, \mathbf{C}_t \rangle + \varepsilon \langle \tilde{\mathbf{P}}_t, \log(\tilde{\mathbf{P}}_t) - 1 \rangle \right) < \sum_{t=1}^s \lambda_t \left(\langle \mathbf{P}_t^\varepsilon, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t^\varepsilon, \log(\mathbf{P}_t^\varepsilon) - 1 \rangle \right).$$

Podemos observar que para cada t, $1 \leq t \leq s$ se tiene la igualdad $\langle \mathbf{P}_t^\varepsilon, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t^\varepsilon, \log(\mathbf{P}_t^\varepsilon) - 1 \rangle = \mathcal{L}^\varepsilon_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t)$ por definición de $\mathbf{P}_t^\varepsilon$. A partir de esta relación se deduce que

$$\sum_{t=1}^s \lambda_t \left(\langle \tilde{\mathbf{P}}_t, \mathbf{C}_t \rangle + \varepsilon \langle \tilde{\mathbf{P}}_t, \log(\tilde{\mathbf{P}}_t) - 1 \rangle \right) < \sum_{t=1}^s \lambda_t \mathcal{L}^\varepsilon_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t). \tag{4.22}$$

Entonces si denotamos por $\tilde{\mathbf{a}} = \tilde{\mathbf{P}}_t \mathbb{1} \in \Sigma_m$, para cada t se tiene que

$$\mathcal{L}^\varepsilon_{\mathbf{C}_t}(\tilde{\mathbf{a}}, \mathbf{b}_t) \leq \langle \tilde{\mathbf{P}}_t, \mathbf{C}_t \rangle + \varepsilon \langle \tilde{\mathbf{P}}_t, \log(\tilde{\mathbf{P}}_t) - 1 \rangle$$

Como consecuencia de la desigualdad anterior y la desigualdad (4.22) se deduce

$$\sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\tilde{\mathbf{a}}, \mathbf{b}_t) < \sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t),$$

de donde se llega a una contradicción con el hecho de que $\mathbf{a}$ es un baricentro regularizado de $\mathbf{b}_1, \ldots, \mathbf{b}_s$ para los pesos dados. □

El recíproco del Lema 4.2.4 también es cierto, es decir, cada solución óptima de (4.21) está asociada a un baricentro regularizado de Wasserstein de $\mathbf{b}_1, \ldots, \mathbf{b}_s$.

Lema 4.2.5. Sean $\mathbf{b}_1, \ldots, \mathbf{b}_s$ vectores de probabilidades tales que $\mathbf{b}_t \in \mathbb{R}^{n_t}$ para cada t, $(\lambda_1, \ldots, \lambda_s) \in \Sigma_s$ un vector de pesos fijado y $\varepsilon > 0$ un factor de regularización. Sean $(\mathbf{P}_t)_t$ matrices de $\mathcal{U}(*, (\mathbf{b}_t)_t)$ en las que se alcanza el mínimo del conjunto (4.21). Entonces el vector de probabilidades $\mathbf{a} = \mathbf{P}_1\mathbb{1} = \cdots = \mathbf{P}_s\mathbb{1} \in \Sigma_m$ es un baricentro regularizado de $\mathbf{b}_1, \ldots, \mathbf{b}_s$ para los pesos dados.

Demostración. Vamos a demostrar el resultado por reducción al absurdo, por ello vamos a asumir que existe un vector $\tilde{\mathbf{a}} \in \Sigma_m$ tal que $\sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\tilde{\mathbf{a}}, \mathbf{b}_t) < \sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t)$. Para cada t consideramos la matriz $\tilde{\mathbf{P}}^{\varepsilon}_t$ en la que se alcanza $\mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\tilde{\mathbf{a}}, \mathbf{b}_t)$. Podemos ver que $(\tilde{\mathbf{P}}^{\varepsilon}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ y por optimalidad de $(\mathbf{P}_t)_t$ se tiene

$$\sum_{t=1}^{s} \lambda_t \left(\langle \mathbf{P}_t, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t, \log(\mathbf{P}_t) - 1 \rangle \right) \leq \sum_{t=1}^{s} \lambda_t \left(\langle \tilde{\mathbf{P}}^{\varepsilon}_t, \mathbf{C}_t \rangle + \varepsilon \langle \tilde{\mathbf{P}}^{\varepsilon}_t, \log(\mathbf{P}^{\varepsilon}_t) - 1 \rangle \right) = \sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\tilde{\mathbf{a}}, \mathbf{b}_t).$$

Además, para cada t podemos asegurar que $\mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t) \leq \langle \mathbf{P}_t, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t, \log(\mathbf{P}_t) - 1 \rangle$, de donde se deduce

$$\sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\mathbf{a}, \mathbf{b}_t) \leq \sum_{t=1}^{s} \lambda_t \left(\langle \mathbf{P}_t, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t, \log(\mathbf{P}_t) - 1 \rangle \right) \leq \sum_{t=1}^{s} \lambda_t \mathcal{L}^{\varepsilon}_{\mathbf{C}_t}(\tilde{\mathbf{a}}, \mathbf{b}_t)$$

en contradicción con lo que se había supuesto. □

Observación 28. Los Lemas 4.2.4 y 4.2.5 indican que existe una correspondencia unívoca entre baricentros regularizados de $\mathbf{b}_1, \ldots, \mathbf{b}_s$ con pesos $\lambda_1, \ldots, \lambda_s$ y matrices $(\mathbf{P}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ en las que se alcanza el mínimo de (4.21). Vamos a emplear esta relación para probar la existencia y unicidad del baricentro regularizado de $\mathbf{b}_1, \ldots, \mathbf{b}_s$.

Proposición 4.2.1 (Existencia y unicidad del baricentro regularizado)**.** *Sean $\mathbf{b}_1, \ldots, \mathbf{b}_s$ vectores de probabilidades tales que $\mathbf{b}_t \in \mathbb{R}^{n_t}$ para cada t, $(\lambda_1, \ldots, \lambda_s) \in \Sigma_s$ un vector de pesos fijado y $\varepsilon > 0$ un factor de regularización. Existe un único baricentro regularizado, $\mathbf{a} \in \Sigma_m$, de $\mathbf{b}_1, \ldots, \mathbf{b}_s$ para los pesos dados.*

Demostración. Atendiendo a la Observación 28, vamos a probar la existencia y unicidad de $(\mathbf{P}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ en las que se alcance el mínimo de (4.21).

En primer lugar, vamos a probar la existencia de un baricentro. Para ello, vamos a demostrar que $\mathcal{U}(*, (\mathbf{b}_t)_t)$ es un conjunto compacto de

$$\mathbb{R}^{m \times n_1} \times \cdots \times \mathbb{R}^{m \times n_s}.$$

Por esta razón, vamos a ver que es un conjunto cerrado y acotado: resulta sencillo comprobar que $\mathcal{U}(*, (\mathbf{b}_t)_t)$ es un conjunto acotado porque todos sus elementos son matrices cuyas componentes son menores que 1; es decir, dado un elemento de $\mathcal{U}(*, (\mathbf{b}_t)_t)$ su norma infinito está acotada por 1. Para demostrar que es cerrado, tomamos una sucesión $\{(\mathbf{P}^l_t)_t\}_{l \geq 1}$ de elementos de $\mathcal{U}(*, (\mathbf{b}_t)_t)$ que converja a $(\mathbf{P}_t)_t$. Vamos a probar que $(\mathbf{P}_t)_t$ también pertenece a $\mathcal{U}(*, (\mathbf{b}_t)_t)$. Fijado un índice t, $1 \leq t \leq s$, se verifica $(\mathbf{P}^l_t)^T\mathbb{1} = \mathbf{b}_t$ para cada $l \geq 1$. En consecuencia, tomando límites en la expresión, por continuidad se deduce

$$(\mathbf{P}_t)^T\mathbb{1} = \lim_{l \to \infty} (\mathbf{P}^l_t)^T\mathbb{1} = \lim_{l \to \infty} \mathbf{b}_t = \mathbf{b}_t.$$

Sea $2 \leq t \leq s$ un índice fijado. Dado que $(\mathbf{P}_t^l)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$, se tiene que

$$\mathbf{P}_t^l \mathbb{1} - \mathbf{P}_1^l \mathbb{1} = (\mathbf{P}_t^l - \mathbf{P}_1^l)\mathbb{1} = \mathbf{0}$$

para cada $l \geq 1$ por definición del conjunto $\mathcal{U}(*, (\mathbf{b}_t)_t)$. Tomando límites en la expresión igualdad anterior se deduce

$$\mathbf{P}_t \mathbb{1} - \mathbf{P}_1 \mathbb{1} = (\mathbf{P}_t - \mathbf{P}_1)\mathbb{1} = \lim_{l \to \infty} (\mathbf{P}_t^l - \mathbf{P}_1^l)\mathbb{1} = \lim_{l \to \infty} \mathbf{0} = \mathbf{0},$$

es decir, $\mathbf{P}_1 \mathbb{1} = \mathbf{P}_t \mathbb{1}$. Por lo tanto, se concluye que $(\mathbf{P}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$.

Además, resulta sencillo ver que este conjunto es no vacío. Para ello tomamos las matrices $(\mathbf{P}_t)_t$ de $\mathbb{R}^{m \times n_1} \times \cdots \times \mathbb{R}^{m \times n_s}$ cuyas componentes vienen dadas por

$$p_{t,i,j} = \frac{1}{m} \cdot b_{t,j},$$

Es decir, $\mathbf{P}_t = diag(\mathbb{1}_m/m) \cdot \mathbb{1}_{m \times n_t} \cdot diag(\mathbf{b}_t)$. Claramente, para cada t se verifica

$$(\mathbf{P}_t)^T \mathbb{1}_m = \mathbf{b}_t \odot \mathbb{1}_{n_t \times m} \cdot \mathbb{1}_m/m = \mathbf{b}_t \odot \mathbb{1}_{n_t} = \mathbf{b}_t$$

y $\mathbf{P}_t \mathbb{1}_{n_t} = \mathbb{1}_m/m \odot \mathbb{1}_{m \times n_t} \cdot \mathbf{b}_t = \mathbb{1}_m/m \odot \mathbb{1}_m = \mathbb{1}_m/m$. Por lo tanto, $(\mathbf{P}_t)_t$ pertenece a $\mathcal{U}(*, (\mathbf{b}_t)_t)$.

Consideramos ahora la función $\mathcal{B}_E((\mathbf{P}_t)_t) = \sum_{t=1}^s \lambda_t \left(\langle \mathbf{P}_t, \mathbf{C}_t \rangle + \varepsilon \langle \mathbf{P}_t, \log(\mathbf{P}_t) - 1 \rangle \right)$ definida en $\mathcal{U}(*, (\mathbf{b}_t)_t)$. Podemos comprobar que se trata de una función continua pues es una combinación lineal de funciones continuas. En virtud del Teorema de Weierstrass, podemos asegurar que existe $(\mathbf{P}_t^0)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ que minimiza $\mathcal{B}_E((\mathbf{P}_t)_t)$.

Para probar la unicidad del óptimo, vamos a emplear un argumento basado en la convexidad, para lo cual vamos a necesitar probar primero la convexidad del conjunto $\mathcal{U}(*, (\mathbf{b}_t)_t)$: tomamos un escalar $\mu \in [0, 1]$ y $(\mathbf{P}_t)_t, (\mathbf{Q}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ entonces para cada t se tiene:

- $\mu \cdot \mathbf{P}_t + (1-\mu) \cdot \mathbf{Q}_t)^T \mathbb{1} = \mu \cdot (\mathbf{P}_t)^T \mathbb{1} + (1-\mu) \cdot (\mathbf{Q}_t)^T \mathbb{1} = \mu\, \mathbf{b}_t + (1-\mu)\, \mathbf{b}_t = \mathbf{b}_t,$
- $[(\mu \cdot \mathbf{P}_t + (1-\mu) \cdot \mathbf{Q}_t) - (\mu \cdot \mathbf{P}_1 + (1-\mu) \cdot \mathbf{Q}_1)]\, \mathbb{1} = \mu \cdot (\mathbf{P}_t - \mathbf{P}_1)\mathbb{1} + (1-\mu) \cdot (\mathbf{Q}_t - \mathbf{Q}_1)\mathbb{1}$
  $= \mu\, \mathbf{0} + (1-\mu)\, \mathbf{0} = \mathbf{0}.$

En consecuencia, $(\mathbf{R}_t)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ donde $\mathbf{R}_t = \mu \cdot \mathbf{P}_t + (1-\mu) \cdot \mathbf{Q}_t$ para cada t, $1 \leq t \leq s$.

Supongamos que existen $(\mathbf{P}_t^0)_t, (\mathbf{Q}_t^0)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ distintos que minimizan $\mathcal{B}_E((\mathbf{P}_t)_t)$ y sea $(\mathbf{R}_t^0)_t \in \mathcal{U}(*, (\mathbf{b}_t)_t)$ dada por

$$\mathbf{R}_t^0 = \tfrac{1}{2} \cdot \mathbf{P}_t^0 + \tfrac{1}{2} \cdot \mathbf{Q}_t.$$

Por la convexidad estricta de $\mathcal{B}_E$, se deduce $\mathcal{B}_E((\mathbf{R}_t^0)_t) = \mathcal{B}_E(\frac{1}{2}(\mathbf{P}_t^0)_t + \frac{1}{2}(\mathbf{Q}_t^0)_t) < \frac{1}{2}\mathcal{B}_E((\mathbf{P}_t^0)_t) + \frac{1}{2}\mathcal{B}_E((\mathbf{Q}_t^0)_t)) = L$, lo que contradice la optimalidad de $(\mathbf{P}_t^0)_t$ y $(\mathbf{Q}_t^0)_t$. □

Los Lemas 4.2.4 y 4.2.5 nos permiten obtener el baricentro de $\mathbf{b}_1, \ldots, \mathbf{b}_s$, el cual es único en virtud de la Proposición 4.2.1, a partir de las matrices $(\mathbf{P}_t^\varepsilon)_t$ en las que se alcanza el mínimo de (4.21): basta con tomar el vector $\mathbf{a} = \mathbf{P}_t^\varepsilon \mathbb{1}$ para cualquier t, ya que estos coinciden. De igual manera, se pueden obtener las matrices $(\mathbf{P}_t^\varepsilon)_t$ en las que se alcanza el mínimo de (4.21) a partir del baricentro $\mathbf{a}$ aplicando el Algoritmo de Sinkhorn. Por esta razón, podemos referirnos a las matrices $(\mathbf{P}_t^\varepsilon)_t$ como matrices óptimas para el problema del baricentro.

Proposición 4.2.2. *Sean $\mathbf{b}_1, \ldots, \mathbf{b}_s$ vectores de probabilidades tales que $\mathbf{b}_t \in \mathbb{R}^{n_t}$ para cada t. Sean $(\lambda_1, \ldots, \lambda_s) \in \Sigma_s$ y $\varepsilon > 0$ un factor de regularización. Entonces las matrices óptimas del problema del baricentro de $\mathbf{b}_1, \ldots, \mathbf{b}_s$ con pesos $\lambda_1, \ldots, \lambda_s$ están dadas por*

$$\mathbf{P}_t^\varepsilon = diag(\mathbf{u}_t) \cdot \mathcal{K}_{t,\varepsilon} \cdot diag(\mathbf{v}_t), \tag{4.23}$$

donde $\mathbf{u}_t \in \mathbb{R}_{>0}^m$ y $\mathbf{v}_t \in \mathbb{R}_{>0}^{n_t}$ para cada t, $1 \leq t \leq s$. Además los vectores $(\mathbf{u}_t)_t$ satisfacen $\prod_{t=1}^s (\mathbf{u}_t)^{\lambda_t} = \mathbb{1}$.

Demostración. En primer lugar, vamos a calcular el Lagrangiano del problema (4.21). Podemos ver que tenemos $m-1$ restricciones de la forma $\mathbf{P}_t\mathbb{1} = \mathbf{P}_1\mathbb{1}$ para cada $t \neq 1$, que se pueden escribir $(\mathbf{P}_i - \mathbf{P}_1)\mathbb{1} = \mathbf{0}$ para cada $i \neq 1$. Para cada una de estas restricciones consideramos el multiplicador de Lagrange f_i. Además, para cada t, $1 \leq t \leq s$, tenemos n_t restricciones dadas por $(\mathbf{P}_t)^T\mathbb{1} = \mathbf{b}_t$: estas son $\sum_{i=1}^{m} p_{t,i,j} = b_{t,j}$ para las cuales consideramos los multiplicadores de Lagrange $g_{t,j}$. Entonces la función Lagrangiana del problema de minimización de (4.21) es:

$$\mathfrak{L}((\mathbf{P}_t)_t, (\mathbf{f}_t), (\mathbf{g}_t)) = \sum_{t=1}^{s} \lambda_t \left(\langle \mathbf{P}_t, \mathbf{C}_t\rangle + \varepsilon\langle \mathbf{P}_t, \log(\mathbf{P}_t) - 1\rangle\right) - \sum_{t=2}^{s}\langle \mathbf{f}_t, (\mathbf{P}_t - \mathbf{P}_1)\mathbb{1}\rangle - \sum_{t=1}^{s}\langle \mathbf{g}_t, (\mathbf{P}_t)^T\mathbb{1} - \mathbf{b}_t\rangle$$

Derivando respecto de cada una de las componentes de $\mathbf{P}_t$ para cada t obtenemos

$$\begin{aligned}\frac{\partial \mathfrak{L}}{\partial p_{t_0,i,j}}((\mathbf{P}_t)_t, (\mathbf{f}_t), (\mathbf{g}_t)) = \lambda_t \left(c_{t_0,i,j} + \varepsilon\left(\log(p_{t_0,i,j}) - 1 + p_{t_0,i,j}\cdot\frac{1}{p_{t_0,i,j}}\right)\right) \\ - \frac{\partial\left(\sum_{t=2}^{s}\langle \mathbf{f}_t, (\mathbf{P}_t - \mathbf{P}_1)\mathbb{1}\rangle\right)}{\partial p_{t_0,i,j}} - g_{t_0,j}\end{aligned} \tag{4.24}$$

Para obtener una expresión simple de $\frac{\partial \mathfrak{L}}{\partial p_{t,i,j}}$, desarrollamos la derivada parcial:

$$\frac{\partial\left(\sum_{t=2}^{s}\langle \mathbf{f}_t, (\mathbf{P}_t - \mathbf{P}_1)\mathbb{1}\rangle\right)}{\partial p_{t_0,i,j}}((\mathbf{P}_t)_t, (\mathbf{f}_t), (\mathbf{g}_t)) = \begin{cases} f_{t_0,i} & \text{si } t_0 \neq 1 \\ -\sum_{t=2}^{s} f_{1,i} & \text{si } t_0 = 1. \end{cases} \tag{4.25}$$

Si definimos $\mathbf{f}_1 = -\sum_{t=2}^{s}\mathbf{f}_t$, podemos escribir la expresión (4.25) de una forma más compacta:

$$\frac{\partial\left(\sum_{t=2}^{s}\langle \mathbf{f}_t, (\mathbf{P}_t - \mathbf{P}_1)\mathbb{1}\rangle\right)}{\partial p_{t_0,i,j}}((\mathbf{P}_t)_t, (\mathbf{f}_t), (\mathbf{g}_t)) = f_{t_0,i}.$$

Además, por definición de $\mathbf{f}_1$ se verifica que $\sum_{t=1}^{s}\mathbf{f}_t = \mathbf{0}$. Ahora, si reescribimos y simplificamos la igualdad (4.24) obtenemos

$$\frac{\partial \mathfrak{L}}{\partial p_{t_0,i,j}}((\mathbf{P}_t)_t, (\mathbf{f}_t), (\mathbf{g}_t)) = \lambda_t\left(c_{t_0,i,j} + \varepsilon\log(p_{t_0,i,j})\right) - f_{t_0,i} - g_{t_0,j}.$$

Las únicas matrices $(\mathbf{P}_t^{\varepsilon})_t$ en las que se alcanza el mínimo de (4.21), deben satisfacer que

$$\frac{\partial \mathfrak{L}}{\partial p_{t_0,i,j}}((\mathbf{P}_t^{\varepsilon})_t, (\mathbf{f}_t), (\mathbf{g}_t)) = 0$$

para cada $1 \leq i \leq m$ y $1 \leq j \leq n_t$ para unos ciertos $(\mathbf{f}_t)_t$ y $(\mathbf{g}_t)_t$. Despejando podemos obtener una expresión para la componente (i,j) de la matriz $\mathbf{P}_t^{\varepsilon}$:

$$p_{t,i,j} = e^{\frac{1}{\varepsilon}(f_{t,i}/\lambda_t - c_{t,i,j} + g_{t,j}/\lambda_t)} = e^{\frac{f_{t,i}}{\lambda_t\,\varepsilon}}\cdot e^{-\frac{c_{t,i,j}}{\varepsilon}}\cdot e^{\frac{g_{t,j}}{\lambda_t\,\varepsilon}}.$$

Denotando por $u_{t,i} = e^{\frac{f_{t,i}}{\lambda_t\,\varepsilon}}$ y $v_{t,j} = e^{\frac{f_{t,i}}{\lambda_t\,\varepsilon}}$ se concluye el primer resultado.

Para probar el segundo resultado, basta comprobar que $(\mathbf{u}_t)^{\lambda_t} = \left(e^{\frac{\mathbf{f}_t}{\lambda_t\,\varepsilon}}\right)^{\lambda_t} = e^{\frac{\mathbf{f}_t}{\varepsilon}}$ y por lo tanto se concluye

$$\prod_{t=1}^{s}(\mathbf{u}_t)^{\lambda_t} = \prod_{t=1}^{s} e^{\frac{\mathbf{f}_t}{\varepsilon}} = e^{\frac{1}{\varepsilon}\sum_{t=1}^{s}\mathbf{f}_t} = e^{\mathbf{0}} = \mathbb{1}.$$

□

La descomposición dada en la Proposición 4.2.2 de las matrices óptimas $(\mathbf{P}_t^\varepsilon)_t$ recuerda a la que se obtuvo para la solución del problema de transporte entrópico discreto en la Proposición 3.0.2. En ese caso se estudió que esta descomposición era esencialmente única en la Observación 18. Esto nos lleva a plantearnos la existencia de algún resultado de unicidad de los vectores $(\mathbf{u}_t)_t$ y $(\mathbf{v}_t)_t$ similar al de los vectores $(\mathbf{u}, \mathbf{v})$ del Algoritmo de Sinkhorn. En la siguiente observación se da solución a esta cuestión y se caracterizan todas las posibles descomposiciones de las matrices óptimas del problema del baricentro.

Observación 29. Podemos notar que si $(\mathbf{u}_t)_t$ y $(\mathbf{v}_t)_t$ son vectores que satisfacen la igualdad (4.23) para cada t, y $(\mu_t)_t$ son escalares estrictamente positivos tales que $\prod_{t=1}^s (\mu_t)^{\lambda_t} = 1$, se tiene que $(\mu_t\, \mathbf{u}_t)_t$ y $\left(\frac{1}{\mu_t}\, \mathbf{v}_t\right)_t$ también verifican (4.23) para cada t. Para probarlo, basta tener en cuenta que si $\mathbf{a}$ es el baricentro regularizado y $(\mathbf{P}_t^\varepsilon)_t$ son las matrices óptimas del problema del baricentro regularizado, entonces para cada t la matriz $\mathbf{P}_t^\varepsilon$ es la solución óptima del problema de transporte entrópico discreto entre $\mathbf{a}$ y $\mathbf{b}_t$. Además, si $(\mathbf{u}'_t)_t$ y $(\mathbf{v}'_t)_t$ satisfacen la condición (4.23), entonces cada par de vectores $(\mathbf{u}'_t, \mathbf{v}'_t)$ debe verificar (3.7) para la matriz $\mathbf{P}_t^\varepsilon$. La Observación 14 permite asegurar que $(\mathbf{u}'_t, \mathbf{v}'_t)$ son de la forma $(\mu_t\, \mathbf{u}_t, \frac{1}{\mu_t}\mathbf{v}_t)$ para un escalar $\mu_t > 0$.

La condición $\prod_{t=1}^s (\mu_t)^{\lambda_t} = 1$ es necesaria, ya que si los vectores $(\mu_t\, \mathbf{u}_t)_t$ y $\left(\frac{1}{\mu_t}\, \mathbf{v}_t\right)_t$ verifican (4.23), se debe cumplir

$$\mathbb{1} = \prod_{t=1}^s (\mu_t\, \mathbf{u}_t)^{\lambda_t} = \left(\prod_{t=1}^s \mu_t^{\lambda_t}\right) \cdot \left(\prod_{t=1}^s \mathbf{u}_t^{\lambda_t}\right) = \left(\prod_{t=1}^s \mu_t^{\lambda_t}\right) \cdot \mathbb{1}.$$

Esta igualdad solo puede darse si $\prod_{t=1}^s (\mu_t)^{\lambda_t} = 1$.

Lema 4.2.6. Sean $\mathbf{b}_1, \ldots, \mathbf{b}_s$ vectores de probabilidades tales que $\mathbf{b}_t \in \mathbb{R}^{n_t}$ para cada t. Sean $(\lambda_1, \ldots, \lambda_s) \in \Sigma_s$ y $\varepsilon > 0$ un factor de regularización. Sean $(\mathbf{u}_t)_t$ y $(\mathbf{v}_t)_t$ tales que las matrices óptimas del problema del baricentro regularizado cumplen $\mathbf{P}_t^\varepsilon = diag(\mathbf{u}_t) \cdot \mathcal{K}_{t,\varepsilon} \cdot diag(\mathbf{v}_t)$ para cada t. Entonces el baricentro regularizado de $\mathbf{b}_1, \ldots, \mathbf{b}_s$ está dado por la expresión

$$\mathbf{a} = \prod_{t=1}^s (\mathcal{K}_{t,\varepsilon} \mathbf{v}_t)^{\lambda_t}.$$

Un razonamiento similar al que se hizo en el Algoritmo de Sinkhorn nos lleva a considerar el Algoritmo 3 para el cálculo del baricentro regularizado en la configuración Euleriana. Este es un algoritmo iterativo y en cada paso se obtienen unas aproximaciones a unos vectores $(\mathbf{u}_t)_t$ y $(\mathbf{v}_t)_t$ que satisfacen (4.23) y una aproximación al baricentro $\mathbf{a}$. La convergencia de este algoritmo está garantizada por los resultado de la Secciones 2.1 y 3.2 de [15].

Configuración Lagrangiana: Al igual que en el problema de proyección, se puede estudiar la versión Lagrangiana del problema del baricentro regularizado. En esta formulación, al igual que en el caso Euleriano, se va a considerar que las probabilidades $\beta_1, \ldots, \beta_s$ tienen soporte finito, tras discretizarlas con alguno de los dos esquemas vistos si fuese necesario. Representaremos la probabilidad β_t, la cual tiene soporte $\{y_{t,1}, \ldots, y_{t,n_t}\}$, por el vector de probabilidades $\mathbf{b}_t \in \Sigma_{n_t}$ como se ha indicado en el Capítulo 3. El enfoque Lagrangiano del problema busca el baricentro de estas probabilidades en el subconjunto de probabilidades con soporte formado por m puntos equiprobables, es decir, probabilidades empíricas de la forma $\alpha = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ para puntos $x_i \in \mathbb{R}^k$. Emplearemos $\mathbf{x}$ para representar la m-tupla $(x_1, \ldots, x_m)$ de puntos de $\mathbb{R}^k$, esto es, $\mathbf{x} \in \mathbb{R}^{m \times k}$. Para cada $1 \leq t \leq s$, la matriz de coste $\mathbf{C}(\mathbf{x})_t$ que consideramos tiene componentes

$$c_{i,j}(\mathbf{x}) = \|x_i - y_{t,j}\|_2^2.$$

Con las notaciones dadas, el problema del baricentro regularizado en la configuración Lagrangiana tiene la siguiente formulación:

$$\arg \min_{\mathbf{x} \in \mathbb{R}^{m \times k}} \sum_{t=1}^s \lambda_t \mathcal{L}^\varepsilon_{\mathbf{C}(\mathbf{x})_t}(\mathbb{1}_m/m, \mathbf{b}_t). \tag{4.26}$$

Algoritmo 3: Algoritmo para el Cálculo de Baricentro

Input: $(\mathbf{b}_t)_t, (\mathbf{C}_t)_t, N$
Output: $(\mathbf{u}_t)_t, (\mathbf{v}_t)_t, (\mathbf{P}_t)_t, \mathbf{a}$
$\mathbf{v}^{(0)} \leftarrow \mathbf{1}$;
while $l \leq N$ **do**
$\mathbf{v}^{(l+1)} \leftarrow \dfrac{\mathbf{b}}{\mathcal{K}_{t,\varepsilon}^T \mathbf{u}^{(l)}}$;
$\mathbf{a}^{(l+1)} \leftarrow \prod_{t=1}^{s} (\mathbf{K}_{t,\varepsilon} \mathbf{v}^{(l+1)})^{\lambda_t}$;
$\mathbf{u}^{(l+1)} \leftarrow \dfrac{\mathbf{a}^{(l+1)}}{\mathbf{K}_{t,\varepsilon} \mathbf{v}_t^{(l+1)}}$;
end
$\mathbf{u}_t \leftarrow \mathbf{u}_t^{(N)}$;
$\mathbf{v}_t \leftarrow \mathbf{v}_t^{(N)}$;
$\mathbf{P}_t \leftarrow diag(\mathbf{u}_t) \cdot \mathbf{K}_{t,\varepsilon} \cdot diag(\mathbf{v}_t)$;
$\mathbf{a} \leftarrow \mathbf{P}_1 \mathbb{1}$

Si empleamos la notación $\mathcal{E}_L(\mathbf{x}; \mathbf{b}_t) = \mathcal{L}^{\varepsilon}_{\mathbf{C}(\mathbf{x})_t}(\mathbb{1}_m/m, \mathbf{b}_t)$ que se introdujo en el problema de proyección Lagrangiano, podemos reescribir (4.26)y obtener una expresión más compacta:

$$\arg \min_{\mathbf{x} \in \mathbb{R}^{m \times k}} \mathcal{B}_E(\mathbf{x}),$$

considerando la función $\mathcal{B}_E(\mathbf{x}) \overset{\text{def.}}{=} \sum_{t=1}^{s} \lambda_t \mathcal{E}_L(\mathbf{x}; \mathbf{b}_t)$.

Observación 30. Al contrario que en la configuración Euleriana, no tenemos garantías de que el problema del baricentro regularizado en la configuración Lagrangiana tenga una solución única.

Como consecuencia de la Observación 25 podemos deducir que la función $\mathcal{E}_L(\mathbf{x}; \mathbf{b}_t)$ es continuamente diferenciable y su gradiente es

$$\nabla \mathcal{E}_L(\mathbf{x}; \mathbf{b}_t) = 2 \left(\frac{1}{m} x_{i,j} - \sum_l p^{\varepsilon}_{t,i,l} \cdot y_{l,j} \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq k}}$$

donde $p^{\varepsilon}_{t,i,j}(\mathbf{x})$ es la componente (i,j) de la matriz de transporte entrópico $\mathbf{P}^{\varepsilon}_t$ entre $\mathbb{1}_m/m$ y $\mathbf{b}_t$ para la matriz de coste $\mathbf{C}(\mathbf{x})$. La regularidad de las funciones $\mathcal{E}_L(\mathbf{x}; \mathbf{b}_t)$, nos permiten afirmar que la función $\mathcal{B}_E(\mathbf{x})$ también es continuamente diferenciable y nos permite dar una expresión del gradiente de $\mathcal{B}_E(\mathbf{x})$ a partir de sus gradientes:

$$\begin{aligned}
\nabla \mathcal{B}_E(\mathbf{x}) &= \sum_{t=1}^{s} \lambda_t \nabla \mathcal{E}_L(\mathbf{x}; \mathbf{b}_t) \\
&= 2 \sum_{t=1}^{s} \lambda_t \left(\frac{1}{m} x_{i,j} - \sum_{l=1}^{n_t} p^{\varepsilon}_{t,i,l} \cdot y_{l,j} \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq k}} \\
&= 2 \left(\frac{1}{m} x_{i,j} - \sum_{t=1}^{s} \left(\lambda_t \sum_{l=1}^{n_t} p^{\varepsilon}_{t,i,l} \cdot y_{l,j} \right) \right)_{\substack{1 \leq i \leq m \\ 1 \leq j \leq k}}
\end{aligned}$$

A partir de esta expresión podemos aplicar el Algoritmo de Descenso de gradiente (Algoritmo 4) para obtener una solución de (4.26).

4.3. Flujos de gradiente

En esta sección se va a estudiar los flujos de gradiente en el espacio de Wasserstein. La motivación del estudio de estos problemas surge de la teoría de flujos de gradiente de $\mathbb{R}^n$. Dada una función diferenciable F, además del estudio del problema de minimización de F en $\mathbb{R}^n$, se puede adoptar un enfoque distinto y estudiar la dinámica que induce F. Es decir, cómo se desplazaría una masa puntual en el espacio si esta se moviese en la dirección de máximo descenso de F. Este concepto se formaliza con los flujos de gradiente, los cuales son curvas diferencibles que son solución de una ecuación diferencial ordinaria. Es posible, rebajar las hipótesis sobre la regularidad de F y estudiar el caso convexo y semiconvexo. El empleo de métodos numéricos para aproximar un flujo de gradiente son los que llevan a considerar el caso extremo en el que no se impone ninguna condición a la función F. Esta perspectiva general conduce a la introducción del método de Movimiento Minimizante que generaliza el esquema implícito de Euler. Este es el punto de partida para definir los flujos de gradiente en espacios métricos generales (ver [2]), entre los cuales se enmarca la teoría de flujos de gradiente en el espacio de Wasserstein. La importancia de estos últimos radica en que permiten analizar el comportamiento de distribuciones de probabilidad respecto del tiempo lo que los hacen especialmente útiles para el estudio de deformaciones.

Sea F una función diferenciable de R^n y sea $x_0 \in \mathbb{R}^n$. Un flujo de gradiente es una curva x que en tiempo 0 vale x_0 y en cada instante se mueve en la dirección de máximo descenso de F, es decir $-\nabla F(x)$. Podemos escribir estas condiciones como un problema de Cauchy:

$$\begin{cases} \dfrac{\partial x}{\partial t} = -\nabla F(x) \\ x(0) = x_0 \end{cases} \tag{4.27}$$

Por un resultado conocido de la teoría de ecuaciones diferenciales ordinarias, podemos asegurar la unicidad de la solución de este problema si imponemos que ∇F sea Lipschitz. El cálculo de esta solución se puede aproximar empleando técnicas de análisis numérico como el esquema de Euler. Este se basa en la discretización del tiempo y en la sustitución de la derivada parcial $\frac{\partial x}{\partial t}$ por una diferencia finita. Es decir, fijamos un paso $\tau > 0$ y consideramos la aproximación

$$\frac{\partial x}{\partial t}(t) \approx \frac{x(t+\tau) - x(t)}{\tau}.$$

Considerando únicamente los puntos de la curva x de la forma $x_k = x(k\tau)$ para $k \geq 0$ y la simplificación anterior obtenemos el método de Euler explícito que viene dado por

$$\begin{cases} \dfrac{x_{k+1} - x_k}{\tau} = -\nabla F(x_k) \\ x_0 = x_0 \end{cases} \tag{4.28}$$

Esta expresión permite obtener la recurrencia $x_{k+1} = x_k - \tau \nabla F(x_k)$, la cual coincide con las iteraciones del Algoritmo de Descenso de gradiente para la longitud de paso fija τ. Existe otro esquema de Euler basado en las mismas suposiciones que en el método explícito. Este emplea el valor del gradiente de F en el punto x_{k+1} en vez de x_k

$$\begin{cases} \dfrac{x_{k+1} - x_k}{\tau} = -\nabla F(x_{k+1}) \\ x_0 = x_0. \end{cases} \tag{4.29}$$

Decimos que (4.29) es el esquema de Euler implícito. Podemos observar que no podemos obtener una expresión explícita para x_{k+1} en función de x_k al contrario que en el esquema numérico previo. Cada iteración del método implícito requiere obtener la solución de

$$x_{k+1} + \tau \nabla F(x_{k+1}) = x_k,$$

la cual generalmente se calcula con otro método numérico por lo que su cálculo es más costoso. Sin embargo, el método de Euler implícito goza de estabilidad numérica al contrario que en el esquema implícito. En el

Apéndice B se ha destacado que el Algoritmo de Descenso de gradiente con longitud de paso uniforme no permite asegurar que la sea un método de descenso, es decir, si τ no es lo suficientemente pequeño puede darse $F(x_{k+1}) > F(x_k)$. Atendiendo a la equivalencia entre el esquema de Euler explícito y el Algoritmo de Descenso de gradiente con longitud de paso uniforme deducimos la inestabilidad de este método numérico. La estabilidad numérica del método implícito se va a comprobar como consecuencia de una equivalencia que se probará mas adelante.

Existe una definición de flujo de gradiente para el caso en que F es convexa en la que no se asume la diferenciabilidad de F. En esta formulación se sustituye el gradiente de F por el subgradiente (ver [2]) y se adaptan los dos esquemas numéricos de Euler de una forma similar. Se puede relajar aún más las hipótesis sobre F asumiendo únicamente semiconvexidad (ver [16]). Referimos a [16] donde se encuentra un análisis de estos dos casos y se prueba la unicidad del flujo de gradiente en ambas situaciones.

Si no imponemos ninguna condición sobre la función F, debemos emplear una formulación alternativa. El siguiente esquema numérico no emplea el gradiente de F y puede ser empleado en esta situación:

$$x_{k+1} \in \arg\min_x \frac{1}{2}\|x - x_k\|^2 + \tau F(x). \tag{4.30}$$

Decimos que (4.31) es el esquema de Movimiento Minimizante. Si se aplica este método numérico a una función F diferenciable entonces recuperamos el esquema de Euler implícito (4.29). Esta afirmación se debe a que la diferenciabilidad de la función F implica que la función $g(x) = \frac{1}{2}\|x - x_k\|^2 + \tau F(x)$ también es diferenciable y en consecuencia el iterante x_{k+1} dado por la expresión (4.31) debe satisfacer la condición necesaria

$$0 = \nabla g(x_{k+1}) = (x_{k+1} - x_k) + \tau \nabla F(x_{k+1}).$$

Esta igualdad es equivalente a que el iterante x_{k+1} del esquema de Movimiento Minimizante sea el iterante $(k+1)$-ésimo del esquema de Euler implícito. Un razonamiento similar, permite recuperar los esquemas implícitos de Euler en los casos en que F es convexa o semiconvexa [16], por lo tanto el esquema de Movimiento Minimizante engloba a todos ellos. Dado un $k \geq 0$, por definición de los iterantes se puede observar que se satisface

$$\frac{1}{2}\|x_{k+1} - x_k\| + \tau F(x_{k+1}) \leq \frac{1}{2}\|x - x_k\| + \tau F(x)$$

para cada x y en particular tomando $x = x_k$ se tiene

$$\frac{1}{2}\|x_{k+1} - x_k\| + \tau F(x_{k+1}) \leq \tau F(x_k).$$

Es decir, se tiene $F(X_{k+1}) \leq F(x_k)$ de donde se deduce la estabilidad numérica de este esquema y en consecuencia se prueba la estabilidad del método de Euler Implícito.

La aplicación del esquema de Movimiento Minimizante a una función F permite construir un par de curvas $x(t), \tilde{x}(t)$ interpolando la sucesión de iterantes x_k obtenida: la primera de ellas es la curva constante a trozos que se obtiene al tomar

$$x(t) = x_k \text{ para } k = \lfloor t/\tau \rfloor.$$

La curva $x(t)$ no es continua por lo general. La curva $\tilde{x}(t)$ se obtiene de la sucesión $\{x_k\}_{k\geq 0}$ mediante interpolación lineal en los intervalos $(\tau k, \tau(k+1))$ de forma que esta sea continua. La expresión para $\tilde{x}(t)$ es

$$x(t) = \begin{cases} x_k & \text{si } t = k \\ x_k + (t - k\tau) \cdot \dfrac{x_{k+1} - x_k}{\tau} & \text{si } t \in (k, k+1). \end{cases}$$

Ambas curvas coinciden en los instantes $k\tau$ para $k \geq 0$ y bajo condiciones generales sobre F, se puede probar la convergencia uniforme de cada una estas curvas a una curva $x(t)$ tal que $x \in L^1(\mathbb{R}^n)$ y $x' \in L^1(\mathbb{R}^n)$ para una sucesión de pasos τ_j que converja a 0. Entonces, si F es de clase $\mathcal{C}^1$ se tiene que la curva límite es solución

(4.27) y por lo tanto es un flujo de gradiente de F. En [16, Prop.2.3] se prueba este resultado y una versión análoga para el caso en que F sea semiconvexa. En esta situación decimos que la curva $x(t)$ son Movimientos Minimizantes Generalizados.

Las propiedades que se han descrito del esquema de Movimiento Minimizante junto con su formulación, en la cual no interviene el gradiente de F ni interviene ninguna derivada parcial, van a motivar su empleo para generalizar los flujos de gradiente en espacios métricos. También se emplean otras caracterizaciones de los flujos de gradiente en $\mathbb{R}^n$ para definir los análogos en espacios métricos. Un caso particular de espacio métrico en el que se puede estudiar los flujos de gradiente es el espacio de Wasserstein, sobre el cual se hará un estudio detallado. Para generalizar la formulación (4.31) a un espacio métrico $(\mathcal{X}, d)$ arbitrario basta sustituir la norma por la distancia d para obtener las iteraciones

$$x_{k+1} \in \arg\min_x \frac{1}{2} d(x, x_k)^2 + \tau F(x). \tag{4.31}$$

En la parte I de [2] se desarrolla la teoría de los flujos de gradiente en espacios métricos, mientras que en la parte II se realiza un análisis detallado en el caso particular de flujos de gradiente en el espacio de Wasserstein. En este trabajo vamos a estudiar estos últimos para lo que emplearemos la distancia p de Wasserstein. Nos centraremos en el caso $p = 2$ porque proporciona mejores propiedades y más similitudes con los flujos de gradiente en $\mathbb{R}^n$. El esquema de Movimiento Minimizante en $\mathcal{W}_2$ toma la siguiente expresión:

$$\alpha_{k+1} = \arg\min_{\alpha \in \mathcal{P}_2(\mathcal{X})} \frac{\mathcal{W}_2^2(\alpha, \alpha_k)}{2} + \tau F(\alpha). \tag{4.32}$$

Las iteraciones que se han definido se denominan método de Jordan-Kinderlehrer-Otto o método JKO.

Como ya se ha hecho en las secciones previas, vamos a analizar una simplificación de este problema en la que consideramos que las probabilidades α_k son discretas. De esta forma, podemos aplicar los dos esquemas de discretización que hemos introducido y estudiar la configuración Euleriana y la Lagrangina. La primera de ellas considerará iterantes α_k con suporte contenido en una malla de puntos fija. La configuración Lagrangiana, por el contrario, tratará de buscar iterantes que sean versiones empíricas, es decir, que sean de la forma $\frac{1}{m}\sum_{i=1}^m \delta_{x_i}$.

Configuración Euleriana: Considerando un conjunto fijo de puntos $\{x_1, \dots, x_m\}$, se puede restringir la búsqueda de los iterantes (4.32) a probabilidades α que tengan su soporte contenido en $\{x_1, \dots, x_m\}$. Representando estas por un vector de probabilidades $\mathbf{a} \in \Sigma_m$ como se ha hecho anteriormente, $\alpha = \sum_{i=1}^m a_i \delta_{x_i}$, se obtiene la formulación Euleriana para el método JKO regularizado:

$$\mathbf{a}_{k+1} = \arg\min_{\mathbf{a} \in \Sigma_m} \frac{1}{2} \mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{a}_k) + \tau F(\mathbf{a}),$$

donde la matriz de coste $\mathbf{C}$ es simétrica y sus componentes son las distancias $c_{i,j} = \|x_i - x_j\|_2^2$.

Es posible escribir cada uno de los pasos del método JKO regularizado Euleriano en función de las matrices $\mathbf{P}_{k+1}^{\varepsilon}$ de transporte entrópico entre $\mathbf{a}_{k+1}$ y $\mathbf{a}_k$. Atendiendo a que para cada $k \geq 0$ se verifica $\mathbf{a}_{k+1} = \mathbf{P}_{k+1}^{\varepsilon}\mathbb{1}$, se puede obtener la expresión

$$\mathbf{P}_{k+1} = \arg\min_{\mathbf{P} \in \mathcal{U}(\,\cdot\,, \mathbf{a}_k)} \langle \mathbf{P}, C\rangle + \varepsilon \langle \mathbf{P}, \log(\mathbf{P}) - \mathbb{1}\rangle + \tau F(\mathbf{P}\mathbb{1}) \quad \text{y} \quad \mathbf{a}_{k+1} = \mathbf{P}_{k+1}\mathbb{1} \tag{4.33}$$

de la configuración Euleriana. El conjunto $\mathcal{U}(\,\cdot\,, \mathbf{a}_k)$ que interviene en la expresión previa hace referencia al conjunto de matrices $\mathbf{P}$ estrictamente positivas de $\mathbb{R}^{m \times n}$ para las cuales $\mathbf{P}^T\mathbb{1} = \mathbf{b}$. Es decir, se corresponde con el conjunto de distribuciones en el espacio producto con segunda marginal dada por el vector de probabilidades $\mathbf{b}$.

Cada iteración de (4.33) se pueden entender como la resolución de un problema de transporte entrópico discreto generalizado (3.18): la función τF va a tomar el papel de la función F en la expresión (3.18). Considerando la función

$$G(y) = \begin{cases} 0 & \text{si } y = \mathbf{b} \\ \infty & \text{si } y \neq \mathbf{a}_k \end{cases}$$

se obtiene la misma formulación que la dada en (4.33). Además, notando que en la iteración k-ésima del método JKO en la configuración Euleriana, la transformada de Legendre de G es

$$G^*(\mathbf{g}) = \sup_{\mathbf{b}} \langle \mathbf{g}, \mathbf{b} \rangle - G(\mathbf{b}) = \langle \mathbf{g}, \mathbf{a}_k \rangle,$$

cada iteración se puede escribir como el problema dual

$$\max_{\mathbf{f} \in \mathbb{R}^m, \mathbf{g} \in \mathbb{R}^n} -F^*(-\mathbf{f}) + \langle \mathbf{g}, \mathbf{a}_k \rangle - \varepsilon \sum_{i,j} e^{\frac{f_i + g_j - c_{i,j}}{\varepsilon}}.$$

Para el cálculo de este máximo se puede implementar el Algoritmo de Sinkhorn generalizado (Algoritmo 2), el cual permite calcular $\mathbf{a}_{k+1}$ a partir del iterante $\mathbf{a}_k$.

Configuración Lagrangiana: La versión Lagrangiana del método JKO regularizado se basa en buscar cada iterante en el espacio de probabilidades empíricas de $\mathbb{R}^k$ con soporte formado por m puntos $\{x_1, \ldots, x_m\}$. Una probabilidad $\mathbf{a}_k$ de estas se escribe como $\frac{1}{m} \sum_{i=1}^m \delta_{x_i(k)}$. Con esta simplificación, y escribiendo $\mathbf{x} = (x_1 \ldots, x_m) \in \mathbb{R}^{m \times k}$ para denotar a un soporte de una probabilidad, la formulación del método JKO es la siguiente:

$$\mathbf{x}(k+1) = (x_1(k), \ldots, x_m(k)) = \arg \min_{\mathbf{x} \in \mathbb{R}^{m \times k}} \frac{1}{2} \mathcal{L}_{\mathbf{C}}(\mathbf{x})^{\varepsilon}(\mathbb{1}/m, \mathbb{1}/m) + \tau F(\mathbf{x}) \quad \text{y} \quad \mathbf{a}_{k+1} = \frac{1}{m} \sum_{i=1}^m \delta_{x_i(k+1)}.$$

La matriz de coste $\mathbf{C}(\mathbf{x})$ de la expresión previa viene dada por las distancias $c_{i,j} = \|x_i - x_j(k)\|_2^2$. La ventaja de esta formulación es que si consideramos una longitud de paso τ pequeña se puede aproximar la distancia de Wasserstein entre $\mathbf{a}_{k+1}$y $\mathbf{a}_k$ por la distancia entre los puntos de su soporte:

$$\mathcal{W}_2^2(\mathbf{a}_{k+1}, \mathbf{a}_k) \approx \|x_1(k+1) - x_1(k)\|_2^2 + \cdots + \|x_m(k+1) - x_m(k)\|_2^2 = \|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2^2.$$

Entonces se puede emplear una aproximación del método JKO (4.32) sin incluir el término de regularización considerando las iteraciones

$$\mathbf{x}(k+1) = (x_1(k), \ldots, x_m(k)) = \arg \min_{\mathbf{x} \in \mathbb{R}^{m \times k}} \frac{1}{2} \|\mathbf{x} - \mathbf{x}(k)\|_2^2 + \tau F(\mathbf{x}) \quad \text{y} \quad \mathbf{a}_{k+1} = \frac{1}{m} \sum_{i=1}^m \delta_{x_i(k+1)}. \tag{4.34}$$

donde consideramos la función definida en $\mathbb{R}^{m \times k}$ dada por $F(\mathbf{x}) = F(\frac{1}{m} \sum_{i=1}^m \delta_{x_i(k+1)})$ abusando de notación. La iteración (4.34) se corresponde con el método de Movimiento Minimizante para F, el cual hemos presentado antes. Si además la función, F es diferenciable entonces se obtiene el esquema implícito de Euler.

4.4. Estimador Mínimo Kantorovich

Esta última sección se va a dedicar a presentar los estimadores Mínimo Kantorovich los cuales surgen como alternativa al estimador Máximo Verosímil. La necesidad del empleo de estimadores distintos del estimador Máximo Verosímil aparece en contextos en los que resulta difícil su cálculo o incluso imposible. La situación que se plantea es la siguiente: se tiene un modelo paramétrico $\theta \mapsto \alpha(\theta) \in \mathcal{P}(\mathcal{X})$ y una serie de observaciones $\{x_1, \ldots, x_n\}$ de una distribución desconocida. De forma que se quiere obtener el elemento que minimice una función de pérdida dada. En el caso del estimador máximo verosímil, la función de pérdida es la probabilidad de obtener las observaciones que se tienen bajo una distribución.

Cuando las distribuciones parametrizadas $\alpha(\theta)$ carecen de función de densidad o la distribución de la que se ha obtenido las realizaciones es singular, no se tienen garantías de una estimación óptima. Por ello resulta razonable el empleo de otra función de pérdida para obtener el estimador. Una situación en la que el empleo del

estimador Máximo Verosímil no es útil se tiene cuando las probabilidades $\alpha(\theta)$ son probabilidades inducidas por una misma probabilidad $\xi \in \mathcal{P}(\mathbb{R}^d)$:

$$\alpha(\theta) = (h_\theta)_{\#}\xi$$

donde $h_\theta : \mathbb{R}^d \to \mathbb{R}^k$ con $d \ll k$. Se dice que estos son modelos generativos para los cuales se han desarrollado varios técnicas para la obtención de estimadores basadas en redes neuronales entre las que destacamos los métodos GAN y VAE.

Una alternativa a estos surge de considerar la distancia de Wasserstein como medida de discrepancia entre las probabilidades $(h_\theta)_{\#}\xi$ y la medida empírica $\frac{1}{n}\sum_{j=1}^{n}\delta_{x_j}$. El valor en el que se minimiza esta distancia es el que proporciona el estimador Mínimo Kantorovich que se define como el coste de transporte óptimo

$$\min_{\theta\in\Theta} \mathcal{W}_2^2\left((h_\theta)_{\#}\xi, \frac{1}{n}\sum_{j=1}^{n}\delta_{x_j}\right). \tag{4.35}$$

La expresión anterior es un caso particular del problema de proyección que se ha estudiado en la Sección 4.2.1, para el cual se han aportado técnicas para obtener aproximaciones a su solución. Estas se basan en el empleo de una versión regularizada de la función de pérdida (4.35), la discretización de las probabilidades $\alpha(\theta)$ y la implementación de métodos de descenso de gradiente para el cálculo del mínimo. Para ello se debe asegurar la regularidad de las parametrizaciones h_θ y aplicar la Proposición 3.3.1. El desarrollo de esta teoría se plantea como una posible vía de investigación futura.

Capítulo 5

Conclusiones

El reciente aumento de aplicaciones del aprendizaje automático ha requerido del estudio de modelos matemáticos cada vez más complejos que emplean cantidades masivas de parámetros para ajustarse a conjuntos de datos masivos. En la mayoría de los casos, estos modelos emplean alguna función de pérdida para evaluar su precisión y poder ajustar los parámetros que lo determinan reduciendo el error cometido. La representación de los datos como distribuciones de probabilidad sugiere emplear una función de pérdida que compare la discrepancia de dos probabilidades. En este sentido encontramos varias alternativas clásicas como la divergencia de Kullback-Leibler o las distancias L^p. Recientemente se ha planteado el empleo de la distancia de Wasserstein para evaluar la diferencia entre dos probabilidades. Esta distancia goza de buenas propiedades y destaca entre las alternativas anteriores porque es capaz de transportar la distancia del espacio subyacente al espacio de probabilidades, de forma que concuerda con la idea física de mover una probabilidad en la otra. En este contexto resulta de gran importancia obtener una forma de cálculo rápida del coste de transporte óptimo, la cual se ajuste a problemas de tamaño grande.

La introducción de la teoría del transporte entrópico permite convertir el problema del transporte óptimo en uno con función objetivo estrictamente convexa lo que supone una ventaja computacional en el caso discreto para el cual existe el Algoritmo de Sinkhorn. Este es un algoritmo de punto fijo que se basa en el escalar una matriz por filas y columnas de forma sucesiva. Su implementación emplea únicamente productos matriciales y divisiones entre vectores componente a componente lo que lo hacen fácilmente paralelizable. En consecuencia, el transporte entrópico discreto es un buen candidato para aproximar el coste del transporte óptimo y por lo tanto una posible función de pérdida para problemas de aprendizaje automático. Podemos encontrar ejemplos de su uso en aplicaciones al tratamiento de imágenes [17] y al análisis de tomografías [18].

En este trabajo se ha introducido el problema de transporte óptimo y se han dado los resultados más importantes que respaldan su utilización como función de pérdida. También, se ha definido el espacio de Wasserstein en el cual se han desarrollado todos los resultados de los dos últimos capítulos. Este es un espacio métrico que hereda las propiedades del espacio subyacente, el cual es $\mathbb{R}^k$ en las aplicaciones que hemos considerado. En el Capítulo 2 se ha desarrollado la teoría del transporte entrópico, para el cual se ha probado la existencia y unicidad del plan de transporte entrópico bajo condiciones generales en la función de coste. Además, se ha demostrado la convergencia del transporte entrópico al transporte óptimo cuando el factor de regularización tiende a cero.

El Capítulo 3 está dedicado al estudio del transporte entrópico en el caso discreto. La formulación discreta permite el empleo del lenguaje matricial y de cálculos matriciales básicos para obtener su solución. El Algoritmo de Sinkhorn es el que proporciona esta solución mediante una serie de iteraciones poco costosas. Este supone una ventaja frente al transporte entrópico discreto el cual es un problema de programación lineal para los que las técnicas de resolución usuales no son lo suficientemente rápidas. En este trabajo se ha estudiado la convergencia del algoritmo a la solución óptima empleando para ello la distancia proyectiva de Hilbert, además

de proporcionarse una interpretación del Algoritmo de Sinkhorn a partir de la formulación dual del problema de transporte entrópico discreto. También, se ha expuesto la generalización del problema de transporte entrópico discreto sobre la cual se ha indicado una vía para la obtención del Algoritmo de Sinkhorn generalizado. Por último, se ha detallado las propiedades de regularidad del coste de transporte entrópico respecto de sus argumentos.

El Capítulo 4 del trabajo se centra en el estudio de varios problemas variacionales definidos en el espacio de Wasserstein. Los dos primeros de ellos responden a una formulación general dada por la minimización de un funcional en el espacio de Wasserstein donde el funcional involucra al menos un término relacionado con un coste de transporte óptimo. En particular, se analizan el problema de proyección de una probabilidad sobre un subconjunto y el problema del baricentro. En [17] se muestran aplicaciones del problema del baricentro para el tratamiento de formas y texturas y en [19] se propone su empleo para técnicas de clustering y de reconstrucción de imágenes. La siguiente cuestión a la que se ha dedicado este capítulo es la de los flujos de gradiente en el espacio de Wasserstein. Estos generalizan los flujos de gradiente en el espacio euclídeo y permiten el estudio de la dinámica de distribuciones sometidos a un potencial. Para su estudio se recorre los conceptos básicos de la teoría de flujos de gradiente en $\mathbb{R}^n$ hasta derivar en el método JKO. En todos los problemas que se han estudiado en el capítulo se ha presentado la versión general y después se han introducido simplificaciones para poder aplicar la teoría que se ha expuesto anteriormente. Es decir, se ha analizado la versión regularizada de estos problemas y se ha trabajado en espacios discretos mediante el uso de discretizaciones. Estas se han estudiado previamente al inicio del capítulo y son esenciales para el tratamiento computacional de estos problemas.

Siguiendo con la línea establecida en el último capítulo se puede profundizar en los problemas variacionales que se han expuesto, añadiendo distintas condiciones con las que se obtengan propiedades interesantes. Por ejemplo, en [20] se estudia una mezcla entre los problemas de proyección y del baricentro: se buscan las coordenadas baricéntricas, es decir unos pesos $(\lambda_1, \ldots, \lambda_s) \in \Sigma_s$, para las cuales el baricentro de s probabilidades fijadas con dichos pesos sea lo más próximo a una probabilidad dada. Una vía de investigación puede ser el uso de estas coordenadas baricénticas como medida de extracción de características para problemas de aprendizaje automático. Otro problema variacional se analiza en [21], el cual es una generalización del problema del baricentro para un grafo. Este se puede tratar con técnicas similares a las empleadas en este trabajo y puede ser empleado para tareas de aprendizaje semisupervisado.

Notación

- $\mathcal{P}(\mathcal{X})$: conjunto de distribuciones de probabilidad en $\mathcal{X}$
- $\Pi(\alpha, \beta)$: conjunto de distribuciones conjuntas con marginales α y β
- $\mathcal{L}_c(\alpha, \beta)$: coste de transporte óptimo entre las probabilidades α y β
- $\mathcal{P}_p(\mathcal{X})$: conjunto de distribuciones de probabilidad en $\mathcal{X}$ con momento de orden p finito
- $\mathcal{W}_p$: distancia p de Wasserstein
- $\frac{dP}{dQ}(x)$: derivada de Radon Nikodym de P respecto de Q
- $\mathrm{KL}(P|R)$: entropía de P relativa a R
- $\mathcal{L}_c^{\varepsilon}(\alpha, \beta)$: coste de transporte entrópico entre las probabilidades α y β
- $\mathcal{W}_p^{\varepsilon}$: distancia p de Wasserstein con factor de regularización ε
- Σ_m: conjunto de vectores de probabilidad de m componentes.
- $\mathcal{U}(\mathbf{a}, \mathbf{b})$: conjunto de matrices no negativas que se corresponden con una distribución conjunta de $\Pi(\alpha, \beta)$
- $\mathcal{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b})$: coste de transporte óptimo discreto entre los vectores de probabilidades $\mathbf{a}$ y $\mathbf{b}$ con matriz de coste $\mathbf{C}$
- $\mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a}, \mathbf{b})$: coste de transporte entrópico discreto entre los vectores de probabilidades $\mathbf{a}$ y $\mathbf{b}$ con matriz de coste $\mathbf{C}$
- π^{ε}: solución del problema de transporte entrópico discreto
- $\mathbf{P}^{\varepsilon}$: matriz asociada a π^{ε}, la solución del problema de transporte entrópico discreto
- $d_{\mathcal{H}}$: distancia proyectiva de Hilbert
- $\mathcal{K}_{\varepsilon}$: matriz de Gibbs con factor de regularización ε.
- $(\mathbf{f}, \mathbf{g})$: potenciales óptimos del problema dual de transporte entrópico discreto
- $\mathbf{P}_{F,G}^{\varepsilon}$: solución del problema del transporte entrópico generalizado
- $\mathcal{U}(*, (\mathbf{b}_t)_t)$: conjunto de matrices factibles para el problema del baricentro regularizado de $\mathbf{b}_1, \ldots, \mathbf{b}_s$

Bibliografía

[1] Gabriel Peyre and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

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

[4] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[5] Juan Antonio Cuesta and Carlos Matrán. Notes on the wasserstein metric in hilbert spaces. *The Annals of Probability*, pages 1264–1276, 1989.

[6] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[7] Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

[8] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.

[9] Edward Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

[10] Joseph E Carroll. Birkhoff's contraction coefficient. *Linear algebra and its applications*, 389:227–234, 2004.

[11] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

[12] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *Electronic Journal of Statistics*, 13(2):5120 – 5150, 2019.

[13] Edouard Pauwels and Samuel Vaiter. The derivatives of sinkhorn–knopp converge. *SIAM Journal on Optimization*, 33(3):1494–1517, 2023.

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

[15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[16] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

[17] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[18] Johan Karlsson and Axel Ringh. Generalized sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4):1935–1962, 2017.

[19] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[20] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

[21] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *International Conference on Machine Learning*, pages 306–314. PMLR, 2014.

[22] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[23] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

[24] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[25] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, third edition, 1986.

[26] Keiô Nagami. Baire sets, borel sets and some typical semi-continuous functions. *Nagoya Mathematical Journal*, 7:85–93, 1954.

[27] torch for R autograd. `https://torch.mlverse.org/technical/autograd/`. Último acceso: 2024-09-19.

[28] PyTorch automatic differentiation package - torch.autograd. `https://pytorch.org/docs/stable/autograd.html`. Último acceso: 2024-09-19.

Apéndice A

Resultados de Teoría de la Probabilidad

En esta sección se van a enunciar una serie de resultados relativos a la convergencia débil de probabilidades en espacios métricos y la derivada de Radon Nikodym, los cuales se emplean a lo largo del trabajo. También se probarán una serie de resultados convergencia en variación total y de aproximación de funciones por sucesiones de funciones continuas y acotadas.

Notación A.1. Sea $(\mathcal{X}, d)$ un espacio métrico. Denotamos por $C_b(\mathcal{X})$ al conjunto de funciones continuas y acotadas de $\mathcal{X}$ en $\mathbb{R}$. Emplearemos $\mathrm{Lip}(\mathcal{X})$ para denotar a las funciones lipschitzianas en $\mathcal{X}$.

Definición A.0.1. Sea $(\mathcal{X}, d)$ un espacio métrico. Sean $\{\mu_n\}_{n\geq 1}$ una sucesión de $\mathrm{Pr}(\mathcal{X})$ y $\mu \in \mathrm{Pr}(\mathcal{X})$. Decimos que la sucesión μ_n converge débilmente a μ si

$$\int_{\mathcal{X}} f(x)\, d\mu_n(x) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f(x)\, d\mu(x)$$

para toda función $f \in C_b(\mathcal{X})$. En esta situación escribimos

$$\mu_n \xrightarrow[n\to\infty]{w} \mu.$$

Teorema A.0.1 (Portmanteau)**.** *Sea $(\mathcal{X}, d)$ un espacio métrico completo y separable. Sean $\{\mu_n\}_{n\geq 1}$ una sucesión de $\mathrm{Pr}(\mathcal{X})$ y $\mu \in \mathrm{Pr}(\mathcal{X})$. Entonces, son equivalentes los siguientes:*

1. $\mu_n \xrightarrow[n\to\infty]{w} \mu$.
2. $\displaystyle\int_{\mathcal{X}} f(x)\, d\mu_n(x) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f(x)\, d\mu(x)$ *para cada $f \in C_b(\mathcal{X})$.*
3. $\displaystyle\int_{\mathcal{X}} f(x)\, d\mu_n(x) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f(x)\, d\mu(x)$ *para cada $f \in C_b(\mathcal{X}) \cap Lip(\mathcal{X})$.*
4. $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$ *para cada G abierto.*
5. $\limsup_{n\to\infty} \mu_n(C) \leq \mu(C)$ *para cada C cerrado.*
6. $\lim_{n\to\infty} \mu_n(A) = \mu(A)$ *para cada conjunto medible A con $\mu(Fr(A)) = 0$.*

Definición A.0.2. Sea $(\mathcal{X}, d)$ un espacio métrico. Sea $\Gamma \subset \mathcal{P}(\mathcal{X})$ una familia de probabilidades. Se dice que Γ es ajustada si para cada $\varepsilon > 0$ existe un compacto K tal que $P(\mathcal{X} \setminus K) < \varepsilon$ para cada $P \in \Gamma$.

Teorema A.0.2 (Prokhorov)**.** *Sea $(\mathcal{X}, d)$ un espacio métrico. Si Γ es una familia de probabilidades ajustada de $(\mathcal{X}, d)$ entonces es relativamente compacta. Si $(\mathcal{X}, d)$ es completo y separable y Γ es relativamente compacta entonces Γ es ajustada.*

Corolario A.0.1. *Sea $(\mathcal{X}, d)$ una espacio métrico completo y separable y sea $\Gamma = \{P_1, \ldots, P_n\}$ una familia finita de probabilidades. Entonces Γ es ajustada.*

Definición A.0.3. Sean P, Q dos probabilidades sobre un espacio medible $\mathcal{X}$. Se dice que P es absolutamente continua respecto de Q si para cada conjunto medible $N \subset \mathcal{X}$ tal que $Q(N) = 0$, se tiene que $P(N) = 0$.

Teorema A.0.3. *Sean P, Q dos probabilidades sobre un espacio polaco $\mathcal{X}$ tales que P es absolutamente continua respecto de Q. Entonces, existe una función f tal que*

$$P(A) = \int_A f(x)\, dQ(x)$$

para cada conjunto medible A de $\mathcal{X}$. Además, f es única salvo en un conjunto de probabilidad Q nula. La función f se denomina derivada de Radon-Nikodym y se denota por

$$\frac{dP}{dQ}(x).$$

Demostración. Ver [25, Thm.32.2] □

Definición A.0.4. Sean P, Q dos probabilidades sobre un espacio medible $\mathcal{X}$. Se dice que P y Q son equivalentes si P es absolutamente continua respecto de Q y Q es absolutamente continua respecto de P y se denota $P \sim Q$.

Proposición A.0.1. *Sea $\mathcal{X}$ un espacio polaco. El espacio de probabilidades $\mathcal{P}(\mathcal{X})$ es completo para la convergencia en variación total.*

Demostración. Sea $\{P_n\}_{n\geq 1}$ una sucesión de Cauchy para la convergencia en variación total. Sea A un subconjutno medible de $\mathcal{X}$. Entonces

$$\tfrac{1}{2}|P_n(A) - P_m(A)| \leq \|P_n - P_m\|_{TV} \xrightarrow[n,m\to\infty]{} 0,$$

es decir, la sucesión de números reales $\{P_n(A)\}_{n\geq 1}$ es de Cauchy y por lo tanto converge a $P(A) \in \mathbb{R}$. Veamos que P define una probabilidad en $\mathcal{X}$. En primer lugar, vemos que $P_n(A) \geq 0$ para cada $n \geq 1$, entonces $P(A) = \lim_{n\to\infty} P_n(A) \geq 0$. Un argumento similar prueba que $P(\mathcal{X}) = 1$: $P_n(\mathcal{X}) = 1$ para cada $n \geq 1$, entonces $P(\mathcal{X}) = \lim_{n\to\infty} P_n(\mathcal{X}) = 1$. Por último, sean $\{A_j\}_{j\geq 1}$ subconjuntos de $\mathcal{X}$ disjuntos 2 a 2. Se tiene que

$$P(\sqcup_{j\geq 1} A_j) = \lim_{n\to\infty} P_n(\sqcup_{j\geq 1} A_j) = \lim_{n\to\infty} \textstyle\sum_{j\geq 1} P_n(A_j) = \sum_{j\geq 1} P(A_j)$$

ya que la serie $\sum_{j\geq 1} P_n(A_j)$ es de números positivos.

Solo falta comprobar que P_n converge a P en variación total. Fijado un $\varepsilon > 0$, existe $n_0 \geq 1$ tal que $\frac{1}{2}|P_n(A) - P_m(A)| \leq \|P_n - P_m\|_{TV} < \varepsilon$ para cada $m, n \geq n_0$ y cada conjunto medible A. Tomando el límite cuando m tiende a infinito en la expresión anterior se deduce que $\frac{1}{2}|P_n(A) - P(A)| \leq \varepsilon$ para cada $A \subset \mathcal{X}$ medible. En consecuencia,

$$\tfrac{1}{2} \sup_{A \text{ medible}} |P_n(A) - P(A)| = \|P_n - P\|_{TV} \leq \varepsilon$$

para cada $n \geq n_0$, de donde se deduce la convergencia de P_n a P en variación total. □

Proposición A.0.2. *Sea $\mathcal{X}$ un espacio polaco y A un subconjunto medible de $\mathcal{X}$. Existe una sucesión creciente de funciones continuas y acotadas que converge puntualmente a la función indicatriz de A.*

Demostración. Vamos a probar el resultado para los abiertos de $\mathcal{X}$ que generan la σ-álgebra de Borel de $\mathcal{X}$. De esta forma, deduciremos el resultado para todos los conjuntos medibles de $\mathcal{X}$.

Sea B un abierto de $\mathcal{X}$. Su complementario es cerrado en $\mathcal{X}$ y por lo tanto la función $d(x, B^c)$ es Lipschitz continua en $\mathcal{X}$. Para cada $n \geq 1$ consideramos al función

$$f_n(x) = \text{mín}\,(1, nd(x, B^c))$$

la cual es continua por ser composición de funciones continuas y está acotada por 1. Podemos observar que si $x \in B$, $d(x, B^c) > 0$ y por lo tanto existe $n_0 \geq 1$ para el cual $n_0 d(x, B^c) \geq 1$ y en consecuencia $f_n(x) = 1$ para cada $n \geq n_0$. Además, se tiene que $f_n(x) = 0$ para todo $n \geq 1$ si $x \notin B$. Es decir, f_n converge puntualmente a la función indicatriz de B. Por último, podemos notar que la sucesión de funciones $\{f_n\}_{n\geq 1}$ que hemos construido es creciente. □

Teorema A.0.4 (Baire)**.** *Sea $\mathcal{X}$ un espacio polaco y sea h una función inferiormente semicontinua positiva. Existe una sucesión creciente de funciones continuas y acotadas que converge puntualmente a h.*

Demostración. El teorema es consecuencia del Corolario 1 de [26], ya que todo espacio polaco satisface el axioma de separación T_4. En este resultado se afirma la existencia de una sucesión creciente de funciones continuas $\{f_n\}_{n\geq 1}$ que converge puntualmente a h. Para completar la prueba del teorema, vamos a construir una sucesión creciente de funciones continuas y acotadas que converja puntualmente a h a partir de $\{f_n\}_{n\geq 1}$. Para ello, fijamos un punto $x_0 \in \mathcal{X}$ y una distancia d compatible con la topología de $\mathcal{X}$ y consideramos las funciones

$$g_n(x) = \text{mín}(1, d(x, B(x_0, n)^c) \cdot \text{máx}(0, f_n(x)).$$

Podemos observar que todas ellas son continuas por ser el producto de funciones continuas. Claramente $0 \geq \text{mín}(1, d(x, B(x_0, n)^c) \geq 1$ y $\text{máx}(0, f_n(x)) \geq 0$, por lo tanto $g_n \geq 0$ para cada $n \geq 1$. Además, tomando límites se tiene

$$\begin{aligned}
&\lim_{n\to\infty} \text{mín}(1, d(x, B(x_0, n)^c) = 1 \\
&\lim_{n\to\infty} \text{máx}(0, f_n(x)) = \text{máx}(0, h(x)) = h(x),
\end{aligned}$$

de donde se deduce que g_n converge puntualmente a h. Por último, se puede ver que $g_n(x) \leq g_{n+1}(x)$ para cada $x \in \mathcal{X}$. □

Apéndice B

Algoritmo de Descenso de gradiente

Esta sección está dedicada a la introducción del Algoritmo de Descenso de gradiente (Algoritmo 4) el cuál es uno de los métodos de descenso más utilizados. En el Capítulo 4 se ha propuesto este algoritmo para el cálculo de las soluciones de varios problemas variacionales en el espacio de Wasserstein.

Los métodos de descenso son métodos iterativos que tratan de obtener el mínimo de una función real F a partir de una sucesión minimizante $\{x_n\}_{n\geq 1}$ que verifique $F(x_n) > F(x_{n+1})$ para cada n. Uno de los tipos de métodos de descenso más simples, la búsqueda lineal, se basa en el esquema

$$x_{n+1} = x_n - \tau_n \Delta_n.$$

El vector Δ_n es la dirección de búsqueda del paso n-ésimo y τ_n es la longitud del paso n-ésimo. Entre los algoritmos que entran dentro del marco de los métodos de descenso de búsqueda lineal se encuentra el algoritmo de descenso de gradiente, el cual vamos a describir más adelante.

Definición B.0.1. Sea $F : D \to \mathbb{R}$ una función definida en un abierto convexo de $\mathbb{R}^d$ y sea $x_0 \in D$. Se dice que $v \in \mathbb{R}^d$ es una dirección de descenso de F en el punto x_0 si existe $\delta > 0$ tal que $f(x_0 + tv) < f(x_0)$ para cada $0 < t \leq \delta$.

Proposición B.0.1. *Sea $F : D \to \mathbb{R}$ una función continuamente diferenciable definida en un abierto convexo de $\mathbb{R}^d$ y sea $x_0 \in D$. Si v satisface $\nabla F(x_0) \cdot v < 0$, entonces es una dirección de descenso para F en x_0.*

Demostración. Sea $v \in \mathbb{R}^d$ tal que $\nabla F(x_0) \cdot v < 0$, por continuidad de $\nabla F(x)$ se deduce que existe $\delta > 0$ tal que $\nabla F(x_0 + tv) \cdot v < 0$ para cada $0 < t \leq \delta$. Entonces, fijado un $0 < t \leq \delta$, por el Teorema del punto medio de Lagrange se tiene

$$F(x_0 + tv) = F(x_0) + t\nabla F(\xi) \cdot v \quad \text{para algún } \xi \text{ del segmento } (x_0, x_0 + tv)$$

y en consecuencia $F(x_0 + tv) < F(x_0)$. Por lo tanto, v es una dirección de descenso de F en x_0. □

Se puede observar que si v es una dirección de descenso para una función F en un punto x_0, entonces para cada $\lambda > 0$ el vector λv también lo es. Por esta razón, se puede limitar la búsqueda de direcciones de descenso a vectores de una norma fija.

Lema B.0.1. Sea $F : D \to \mathbb{R}$ una función continuamente diferenciable definida en un abierto convexo de $\mathbb{R}^d$ y sea $x_0 \in D$ en el que no se anula $\nabla F(x)$. El vector $-\nabla F(x_0)^T$ es una dirección de descenso de F en x_0. Además, es el único vector de entre aquellos que tienen norma $\|\nabla F(x_0)\|_2$ en el que se alcanza el mínimo

$$\min_{\|v\|_2 = \|\nabla F(x_0)\|_2} \nabla F(x_0) \cdot v. \tag{B.1}$$

Demostración. El vector $-\nabla F(x_0)^T$ es una dirección de descenso de F en x_0 ya que claramente verifica la condición de la Proposición B.0.1:

$$\nabla F(x_0) \cdot (-\nabla F(x_0)^T) = -\|\nabla F(x_0)\|_2^2 < 0.$$

Para probar la segunda afirmación, empleamos la desigualdad de Cauchy-Schwarz de donde se deduce que $|\nabla F(x_0) \cdot v| \leq \|\nabla F(x_0)\|_2 \cdot \|v\|_2 = \|F(x_0)\|_2^2$. Es decir, $-\|F(x_0)\|_2^2 \leq \nabla F(x_0) \cdot v \leq \|F(x_0)\|_2^2$. La desigualdad inferior se alcanza en $v = -\nabla F(x_0)$ por lo que se tiene que este vector minimiza (B.1). Además, para cualquier otro vector v con norma $\|\nabla F(x_0)\|_2$ que no sea múltiplo de $\nabla F(x_0)$, se da la desigualdad estricta en la desigualdad de Cauchy-Schwarz, por lo que no minimiza (B.1). □

El Lema B.0.1 nos permite considerar $\nabla F(x_0)$ como dirección de descenso de F en el punto x_0 siempre que este sea un vector no nulo. En caso de que $\nabla F(x_0)$ fuese nulo se tendría que x_n es un extremo absoluto de F, un extremo local de F, o bien un punto de silla. Atendiendo a las propiedades que hemos estudiado, vamos a estudiar el método de descenso que se basa en emplear $\nabla F(x_n)$ como dirección de búsqueda del paso n-ésimo.

Sea $F : D \to \mathbb{R}$ una función continuamente diferenciable definida en un abierto convexo de $\mathbb{R}^d$ que tiene un mínimo x^* en D. Una condición necesaria para que x^* sea mínimo es que $\nabla F(x^*) = \mathbf{0}$. Fijada una tolerancia, tol, por la continuidad de $\nabla F(x)$ se deduce que existe una bola $B(x^*, \delta)$ para la cual $\|\nabla F(x)\|_2 <$ tol para cada $x \in B(x^*, \delta)$. Por lo tanto, un algoritmo que implemente el método de descenso para obtener el mínimo de F y emplee $\nabla F(x_n)$ como dirección de búsqueda puede emplear $\|\nabla F(x_n)\|_2 <$ tol como condición de parada.

Teniendo presente, las observaciones que hemos realizado, resulta natural considerar el Algoritmo 4 para obtener el mínimo de una función en las condiciones dadas. Decimos que este es el Algoritmo de Descenso de gradiente. Podemos ver que no se ha especificado el valor de la longitud del paso n-ésimo y es que existen de diversas maneras de hacerlo. La forma más simple consiste en tomar un valor constante $\tau_n = \tau$ para la longitud del paso. El problema de esta elección es que no existen garantías de que el algoritmo resultante sea un método de descenso, es decir, no se puede asegurar $f(x_{n+1}) < f(x_n)$. Existen otras alternativas para obtener τ_n, entre ellas destacamos el método de búsqueda exacta, el cual se basa en tomar como τ_n el $\tau > 0$ que minimice $F(x_n + \tau \nabla F(x_n))$, o la búsqueda *Backtracking* que obtiene una aproximación al valor dado por la búsqueda exacta. En la Sección 9.2 de [11] se puede encontrar un desarrollo de las propiedades que aportan cada uno de estos métodos.

Algoritmo 4: Algoritmo de Descenso de gradiente

Input: F, x_0, tol
Output: x^*
while $\|\nabla F(x_n)\|_2 \geq$ *tol* **do**
    Calcular $\nabla F(x_n)$;
    Calcular longitud del paso τ_n;
    $x_{n+1} \leftarrow x_n - \tau_n \nabla F(x_n)$;
end
$x^* \leftarrow x_n$;

Apéndice C

Diferenciación Automática

El objetivo de los algoritmos de Diferenciación Automática es obtener la matriz Jacobiana de una aplicación diferenciable $F : \mathbb{R}^{N_0} \to \mathbb{R}^{N_p}$ en un punto $x_0 \in \mathbb{R}^{N_0}$. Estos se basan en la suposición de que F puede escribirse como composición de p aplicaciones diferenciables elementales $f_i : \mathbb{R}^{N_{i-1}} \to \mathbb{R}^{N_i}$ de la siguiente manera:

$$F = f_p \circ f_{p-1} \circ \cdots \circ f_2 \circ f_1.$$

El cálculo de la matriz $\frac{dF}{d\mathbf{x}_0}(x_0)$ se va a realizar de forma recursiva, para ello se va a hacer uso de la Regla de la cadena. En primer lugar, para cada $1 \leq k \leq p-1$ definimos las aplicaciones

$$F_k = f_k \circ f_{k-1} \circ \cdots \circ f_2 \circ f_1$$
$$F^k = f_p \circ f_{p-1} \circ \cdots \circ f_{k+2} \circ f_{k+1}.$$

Consideraremos $F_0 = Id_{\mathbb{R}^{N_0}}$ y $F^p = Id_{\mathbb{R}^{N_p}}$. De esta forma, podemos observar que se verifica $F^k \circ F_k = F$ para cada $0 \leq k \leq p$. Además, se satisfacen las siguientes relaciones:

$$\begin{aligned} F_k &= f_k \circ F_{k-1} \qquad \text{para cada } 1 \leq k \leq p, \\ F^k &= F^{k+1} \circ f_{k+1} \quad \text{para cada } 0 \leq k \leq p-1. \end{aligned} \tag{C.1}$$

Para cada $0 \leq k \leq p$ escribiremos $x_k = F_k(x_0) \in \mathbb{R}^{N_k}$, entonces por (C.1) se verifica $x_k = f^k(x_{k-1})$ para cada $1 \leq k \leq p$. También, vamos a emplear la notación $\mathbf{x}_k$ para referirnos a las variables de una aplicación con salida $\mathbb{R}^{N_k}$. Por esta razón escribiremos

$$\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$$

para designar a la matriz Jacobiana de f_k en el punto x_{k-1}. De acuerdo con esta notación, utilizaremos

$$\frac{dF_k}{d\mathbf{x}_0}(x_0) \quad \text{y} \quad \frac{dF^k}{d\mathbf{x}_k}(x_k)$$

para referirnos a la matriz Jacobiana de F_k en x_0 y a la matriz Jacobiana de F^k en x_k respectivamente. Como consecuencia de las igualdades (C.1) se obtienen las siguientes expresiones aplicando la Regla de la cadena.

$$\frac{dF_k}{d\mathbf{x}_0}(x_0) = \left[\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})\right] \cdot \left[\frac{dF_{k-1}}{d\mathbf{x}_0}(x_0)\right] \qquad \text{para cada } 1 \leq k \leq p, \tag{C.2}$$

$$\frac{dF^k}{d\mathbf{x}_k}(x_k) = \left[\frac{dF^{k+1}}{d\mathbf{x}_{k+1}}(x_{k+1})\right] \cdot \left[\frac{df_{k+1}}{d\mathbf{x}_k}(x_k)\right] \qquad \text{para cada } 0 \leq k \leq p-1. \tag{C.3}$$

Atendiendo a que $F_p = F$, podemos calcular la matriz Jacobiana de F en x_0 de forma recursiva mediante la igualdad (C.2). En un bucle se calculan x_k y las matrices $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$ de forma sucesiva y se actualiza

Algoritmo 5: Algoritmo de Diferenciación Automática Forward

Input: $x_0, f_1, f_2, \ldots, f_{p-1}, f_p$
Output: $\frac{dF}{d\mathbf{x}_0}(x_0)$
$\frac{dF_0}{d\mathbf{x}_0}(x_0) \leftarrow Id$;
for $1 \leq k \leq p$ **do**
  Calcular $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$;
  $\frac{dF_k}{d\mathbf{x}_0}(x_0) \leftarrow \left[\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})\right] \cdot \left[\frac{dF_{k-1}}{d\mathbf{x}_0}(x_0)\right]$;
  $x_k \leftarrow f_k(x_{k-1})$;
end
$\frac{dF}{d\mathbf{x}_0}(x_0) \leftarrow \frac{dF_p}{d\mathbf{x}_0}(x_0)$;

el valor $\frac{dF_k}{d\mathbf{x}_0}(x_0)$ hasta obtener $\frac{dF_p}{d\mathbf{x}_0}(x_0) = \frac{dF}{d\mathbf{x}_0}(x_0)$. El procedimiento resultante se denomina Algoritmo de Diferenciación Automática Forward y se muestra en el Algoritmo 5.

De una forma similar se puede emplear la igualdad (C.3) de forma recursiva para calcular la matriz Jacobiana de F notando que $F^0 = F$. Para ello, primero se calculan los valores x_k y las matrices $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$ las cuales se almacenan en memoria. En un segundo bucle se realizan los productos

$$\left[\frac{dF^{k+1}}{d\mathbf{x}_{k+1}}(x_{k+1})\right] \cdot \left[\frac{df_{k+1}}{d\mathbf{x}_k}(x_k)\right]$$

hasta obtener la matriz Jacobiana de F en x_0. De esta forma se obtiene el Algoritmo 6, el cual se denomina Algoritmo de Diferenciación Automática Backward.

Algoritmo 6: Algoritmo de Diferenciación Automática Backward

Input: $x_0, f_1, f_2, \ldots, f_{p-1}, f_p$
Output: $\frac{dF}{d\mathbf{x}_0}(x_0)$
for $1 \leq k \leq p-1$ **do**
  Calcular $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$;
  $x_k \leftarrow f_k(x_{k-1})$;
end
$\frac{dF^p}{d\mathbf{x}_0}(x_0) \leftarrow Id$;
for $0 \leq k \leq p-1$ **do**
  $\frac{dF^k}{d\mathbf{x}_k}(x_k) \leftarrow \left[\frac{dF^{k+1}}{d\mathbf{x}_{k+1}}(x_{k+1})\right] \cdot \left[\frac{df_{k+1}}{d\mathbf{x}_k}(x_k)\right]$;
end
$\frac{dF}{d\mathbf{x}_0}(x_0) \leftarrow \frac{dF^0}{d\mathbf{x}_0}(x_0)$;

La diferencia entre los dos algoritmos se encuentra en el orden en el que se multiplican las matrices Jacobianas intermedias $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$: el método Forward realiza los productos en orden ascendente de k, mientras que el método Backward realiza los productos en orden descendente de k. La principal ventaja del método Backward frente al método Forward se encuentra cuando N_p es pequeño en comparación con N_0. Esto se debe a que el número de productos necesarios para el cálculo del producto (C.2) es $N_k \times N_{k-1} \times N_0$ y por lo tanto el Algoritmo 5 realiza

$$\sum_{k=1}^{p}(N_k \times N_{k-1} \times N_0) = N_0 \sum_{k=1}^{p}(N_{k-1} \times N_k)$$

productos (sin contar el cálculo de $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$). El cálculo de los productos (C.3) requiere de $N_p \times N_{k+1} \times N_k$

productos que suman un total de

$$\sum_{k=0}^{p-1}(N_p \times N_{k+1} \times N_k) = N_p \sum_{k=0}^{p-1}(N_k \times N_{k+1}) = N_p \sum_{k=1}^{p}(N_{k-1} \times N_k)$$

productos (sin contar el cálculo de $\frac{df_k}{d\mathbf{x}_{k-1}}(x_{k-1})$). En consecuencia, si $N_p \ll N_0$, el número de productos que realiza en método Backward es significativamente menor que el método Forward. A cambio, el Algoritmo de Diferenciación Automática Backward necesita guardar los resultados intermedios del primer bucle.

En la práctica la implementación de estos algoritmos es distinta a la que hemos descrito ya que raramente se necesita la matriz Jacobiana de F entera sino su producto por un vector. La implementación se basa en la construcción de un grafo dirigido que indica las dependencias entre cada variable. Este se recorre hacia detrás para calcular las derivadas parciales mediante el método Backward. La diferenciación automática está implementada en muchos lenguajes de programación, por ejemplo en la librería torch de R y en la librería PyTorch de Python, ver [27] y [28].