UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

BARICENTROS DÉBILES DE WASSERSTEIN: TEORÍA Y APLICACIONES EN APRENDIZAJE DE MÁQUINAS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

TOMÁS IGNACIO VALENCIA DROGUETT

PROFESOR GUÍA:
Felipe Tobar Henríquez

MIEMBROS DE LA COMISIÓN:
Joaquín Fontbona Torres
Elsa Cazelles

Este trabajo ha sido parcialmente financiado por:
CMM ANID BASAL FB210005.

SANTIAGO DE CHILE
2024

RESUMEN DE TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS Y MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO
POR: TOMÁS IGNACIO VALENCIA DROGUETT
FECHA: 2024
PROF. GUÍA: FELIPE TOBAR

BARICENTROS DÉBILES DE WASSERSTEIN: TEORÍA Y APLICACIONES EN APRENDIZAJE DE MÁQUINAS

En la matemática existe el área de transporte óptimo, la cual durante los últimos años ha recibido bastante atención principalmente por parte de la comunidad de aprendizaje de máquinas. Una de las herramientas que permite definir esta teoría es la noción de baricentro de Wasserstein [1], que es la extensión natural del promedio de puntos al espacio de distribuciones de probabilidad. Recientemente, se propuso la teoría del transporte óptimo débil, la cual consiste en una generalización del transporte óptimo [2]. Con base en esta formulación, en [3] se introducen los baricentros débiles de una familia de distribuciones de probabilidad. Se proporcionó un análisis teórico de este objeto, se discutió su interpretación a la luz del orden convexo entre medidas de probabilidad y también se presentó un algoritmo iterativo para calcular un baricentro débil para una familia finita de distribuciones de entrada. Sin embargo, este baricentro es difícil de computar y no existen otras aproximaciones para este cálculo, además de que la formulación matemática del problema de transporte óptimo débil no presenta simetría, a diferencia del transporte óptimo clásico. El presente estudio diseña un algoritmo eficiente para resolver el problema de baricentro débil en alta dimensionalidad, junto con aprovechar la asimetría mencionada para introducir la noción de baricentro débil reverso y extraer propiedades matemáticas de este objeto. En particular, se muestra que a diferencia del baricentro débil, el cual extrae información geométrica común compartida por todas las distribuciones de entrada, codificada como una variable aleatoria latente que las subyace a todas ellas. El baricentro débil reverso posee la propiedad de que extrae toda la información de las distribuciones de entrada, y genera que, bajo ciertas condiciones, sean variable latente para el. También se diseñó un algoritmo para computar este último baricentro en el caso particular de distribuciones gaussianas unidimensionales. Además, se calcularon los respectivos baricentros sobre distribuciones gaussianas unidimensionales y se realizaron diferentes comparaciones entre estos tres baricentros, donde se ilustran los aspectos teóricos de cada uno. Finalmente, se calculó el baricentro débil en el dataset MNIST de alta dimensionalidad y se realizaron comparaciones entre este baricentro y el de Wasserstein. En todos los experimentos se da cuenta de las propiedades de variable latente que poseen estos baricentros débiles.

Nada es verdad, todo está permitido
Saludos

Agradecimientos

En primer lugar quiero agradecer a mis abuelos, a Raulito y Monya Vargas. Quienes me han me han dado un apoyo y cariño incondicional que lo llevaré siempre conmigo. Esta tesis es para ustedes. También agradezco a mis padres por su cariño y apoyo.

Agradezco a mis amigos del colegio, en particular al Chico, Camilo y Cristobal. Muy buenos recuerdos del colegio me llevo con ustedes. También a mis amigos de plan común, sobre todo al Jaimeme, Mario y Juan, son unos cracks y la he pasado muy bien con ustedes. Una mención especial a mi padre Ken Miyake, que lo conozco desde primer año y se volvió un amigo muy cercano y muy pana.

Agradezco a mis amigos de especialidad, a todes LES OTRES C.S en verdad son unos genios, muy buenas personas y estoy feliz de haberlos conocido y que formen parte de mi vida. En particular a la Vale, Clemun, Yeni, Banduc, Branco, Isaac, Kun, Laengle que tiene una sabiduría infinita, tremendos consejos, apoyo y risas. Tapia que hemos compartido bastante desde plan común, y también muy buenos momentos, conversaciones, risas. Panchito por ser tan lindo, crack, apañador. JP quien estuvo en momentos de cambios importantes en mi vida, por tus buenos consejos, risas, apoyo. y por supuesto que siempre estuvieron dispuestos a ayudarme cuando lo necesite. También agradezo a los cracks de AMOGUS.

Agradezco a la gente de ISATEC, en verdad ha sido un grupo de trabajo muy bueno. En particular al Xavi, Matu y Axel, que hemos compartido bastante este último año y me he reido mucho con ustedes. Sobre todo con el Axel que hemos compartido bastante en poco tiempo.

También agradezco a mi profesor guía Felipe Tobar por su apoyo, disponibilidad y preocupación por el desarrollo de esta tesis. Y sobre todo por las buenas conversaciones en las reuniones después de hablar de la tesis. También agradezco al profesor Joaquín Fontbona por su apoyo en el desarrollo de esta tesis y por ser parte de la comisión. Finalmente agradezco a Elsa Cazelles por sus comentarios y disponibilidad en ser parte de la comisión de esta tesis. Agradezco a toda la gente que a pesar de no estar conmigo ahora fueron importantes en mi vida.

Mencion honrosa a la mismisima luna, que me alegra la existencia con sus maullidos sobre todo al recibirme cada día. Finalmente agradezco a mi mentor Jacobo.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Antes de introducir la problemática, se darán algunas notaciones y definiciones esenciales del área de transporte óptimo en general. Posteriormente, se detallará la motivación y problema a investigar en este trabajo.

1.1. Notaciones

En esta sección se insertarán las notaciones que se utilizarán a lo largo del documento.

Notación 1.1.1. Se denotará por $\mathcal{P}(\mathcal{X})$ al *conjunto de medidas de probabilidad* sobre $\mathcal{X}$, dotado de la topología débil. $\mathcal{P}_{ac}(\mathcal{X})$ al *subconjunto de las medidas de probabilidad absolutamente continuas* con respecto a una medida de referencia común σ-finita λ sobre $\mathcal{X}$. $\mathcal{P}_p$ corresponde al *conjunto de las medidas de probabilidad con momento de orden p finito. Convergencia débil* $\mu_k \to \mu$ significa que μ_k converge débil a μ, vale decir $\int \phi d\mu_k \to \int \phi d\mu$ para toda función ϕ continua acotada.

Dado que las medidas discretas son útiles para aterrizar las ideas e ilustrar conceptos en el documento, se definirá una forma estándar de denotarlas. Para esto, primero se revisará el concepto de simplex.

Definición 1.1 (Simplex) *El simplex de probabilidad se define por medio de la siguiente relación:*

$$\Sigma_n \stackrel{\text{def}}{=} \left\{ \mathbf{a} \in \mathbb{R}_+^n \ : \ \sum_{i=1}^n \mathbf{a}_i = 1 \right\} \tag{1.1}$$

El simplex de probabilidad con n bins. También se le suele decir como el conjunto de vectores de probabilidad sobre $\mathbb{R}_+^n$ *donde a los elementos del simplex se les denominará de forma indistinta como histogramas o vectores de probabilidad.*

Notación 1.1.2 (Medida discreta). Una *medida discreta* de pesos $\mathbf{a} \in \mathbb{R}^n$ en las posiciones $x_1, \ldots, x_n \in \mathcal{X}$ se escribirá por

$$\mu = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}. \tag{1.2}$$

donde δ_x es la medida de Dirac concentrado en el punto $x \in \mathcal{X}$, la cual se define por:

$$\delta_x(A) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases} \qquad \forall A \text{ medible.}$$

Si $\mathbf{a} \in \Sigma_n$, entonces $\mu \in \mathcal{P}(\mathcal{X})$.

Notación 1.1.3. Para un elemento $\mu \in \mathcal{P}_{ac}(\mathcal{X})$ se denotará su densidad con respecto a la medida de referencia λ por $\rho_\mu(x) = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}(x)$.

Observación 1.1.4. Si se considera la medida de referencia λ como la cuenta-puntos, entonces para una medida discreta $\mu = \sum_i \mathbf{a}_i \delta_{x_i} \in \mathcal{P}_{ac}(\mathcal{X})$ se tiene que

$$\rho_\mu(x) = \sum_{i \in I(x)} \mathbf{a}_i, \text{ donde } I(x) = \{i \in [n] \mid x = x_i\}$$

con $[n] = \{1, \ldots, n\}$. En la siguiente Figura se ven ejemplos de medidas empíricas y densidades en 1 y 2 dimensiones.

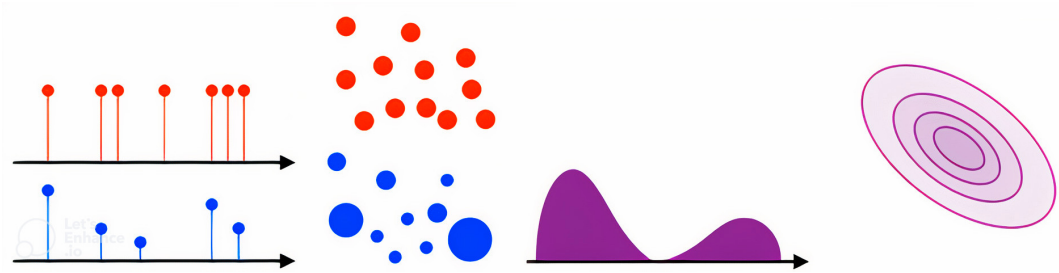


Figura 1.1: Representación esquemática de distribuciones discretas $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ (rojo corresponde a una distribución uniforme empírica con $a_i = 1/n$, y azul a distribuciones arbitrarias) y densidades ρ_μ (en color morado), tanto en una como en dos dimensiones. Las distribuciones discretas en una dimensión se muestran como gráficos de tallos (con longitud igual a a_i), y en dos dimensiones mediante nubes de puntos (imagen obtenida de [4]).

Definición 1.2 (Variable aleatoria latente) *[5] Una variable latente es una variable que no se observa directamente y se asume que afecta a las variables que son observables. Estas variables se usan principalmente para representar el efecto de factores no observados. En mediciones, estas variables representan el resultado verdadero, mientras que las variables observadas son perturbaciones de ellas.*

1.2. Transporte Óptimo

En la matemática existe el área de transporte óptimo (OT, por sus siglas en inglés), la cual ha recibido especial atención recientemente por parte de la comunidad de aprendizaje de máquinas, pues se le ha encontrado diversas aplicaciones tales como *Domain Adaptation* [6]. También, como se verá más adelante, permite definir formalmente una noción de distancia en el espacio de distribuciones de probabilidad conocida como distancia de Wasserstein. Este objeto matemático es importante, ya que en aplicaciones ha permitido obtener mejores resultados, como por ejemplo en [7], donde se propone una alternativa a las GAN, cuya principal diferencia es cambiar la *cross entropy loss* por la distancia de Wasserstein entre distribuciones. A continuación se detallarán los aspectos relevantes de esta teoría con el fin de dar el suficiente contexto en este trabajo.

1.2.1. Formulación de Monge

El problema original propuesto por Monge buscaba transportar de manera óptima una cantidad de arena a un hueco del mismo volumen. Para ilustrar el problema, μ representa una pila de arena, mientras que ν corresponde a un hoyo donde se busca mover la arena. Con el fin de formalizar esta idea, se dará la siguiente definición.

Definición 1.3 (Push-forward) *Para una función medible $T : \mathcal{X} \to \mathcal{Y}$, el operador push-forward $T_{\#} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ es definido por medio de la siguiente relación*

$$T_{\#}\mu(B) = \mu \circ T^{-1}(B) = \mu(T^{-1}(B)), \quad \forall B \subseteq \mathcal{Y} \text{ medible.} \tag{1.3}$$

para algún $\mu \in \mathcal{P}(\mathcal{X})$. De forma equivalente, el operador push-forward se puede definir como aquel operador que satisface la siguiente relación:

$$\forall h \in \mathcal{C}(\mathcal{Y}), \ \int_{\mathcal{Y}} h(y)\, \mathrm{d}(T_{\#}\mu)(y) = \int_{\mathcal{X}} h(T(x))\, \mathrm{d}\mu(x) \tag{1.4}$$

De donde $\mathcal{C}(\mathcal{Y})$ denota a todas las funciones continuas

Observación 1.2.1. Para el caso en que $\mu = \sum_i \mathbf{a}_i \delta_{x_i}$ sea una medida discreta, entonces el operador $T_{\#}$ lo que hará será cambiar las posiciones de las masas, de forma que se cumpla lo siguiente:

$$T_{\#}\mu = \sum_i \mathbf{a}_i \delta_{T(x_i)} \tag{1.5}$$

.

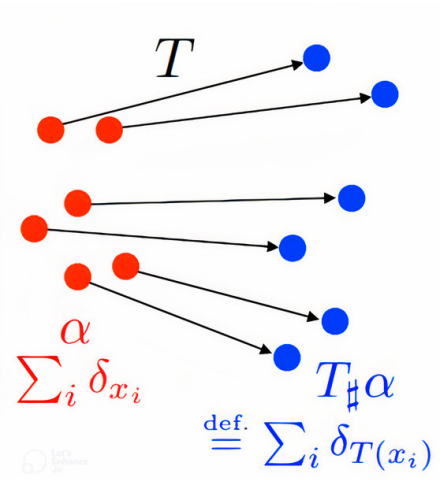


Figura 1.2: Ejemplo de un operador push-forward para una medida α (imagen obtenida de [4])

Intuitivamente, tal como se muestra en la Figura 1.2, el operador *push-forward* $T_{\#}$ es una extensión de la función T, que lo que hace es **mover** una medida de probabilidad sobre el espacio $\mathcal{X}$ a una **nueva** medida de probabilidad en el espacio $\mathcal{Y}$.

Formalmente, se tendrán 2 medidas sobre los espacios $\mathcal{X}$ e $\mathcal{Y}$ respectivamente. Sea $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ una función de costo, donde $c(x, y)$ cuantifica que tanto cuesta mover una

unidad de masa desde $x \in \mathcal{X}$ hasta $y \in \mathcal{Y}$. El problema de transporte óptimo consiste en cómo transportar la medida μ hasta ν minimizando el costo asociado.

Definición 1.4 (Problema de Monge) *: Dadas 2 medidas μ, ν, una función de coste $c(x, y) = \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ (que se le llamará también como coste fuerte) medible y acotada por abajo. El problema general de Monge consiste en encontrar una función T que minimice*

$$M(T) = \inf_T \left\{ \int_{\mathcal{X}} c(x, T(x)) \, \mathrm{d}\mu(x) \mid T_{\#}\mu = \nu \right\} \tag{1.6}$$

En el caso en que las medidas ($\mu = \sum_i \mathbf{a}_i \delta_{x_i}$, $\nu = \sum_j \mathbf{b}_j \delta_{y_j}$) sean discretas, el problema de transporte se reduce a

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \mid T_{\#}\mu = \nu \right\} \tag{1.7}$$

*A la función T que resuelva el problema se le llamará **monge map** o **transporte**. En la imagen 1.3 se ejemplifican estos mapeos.*

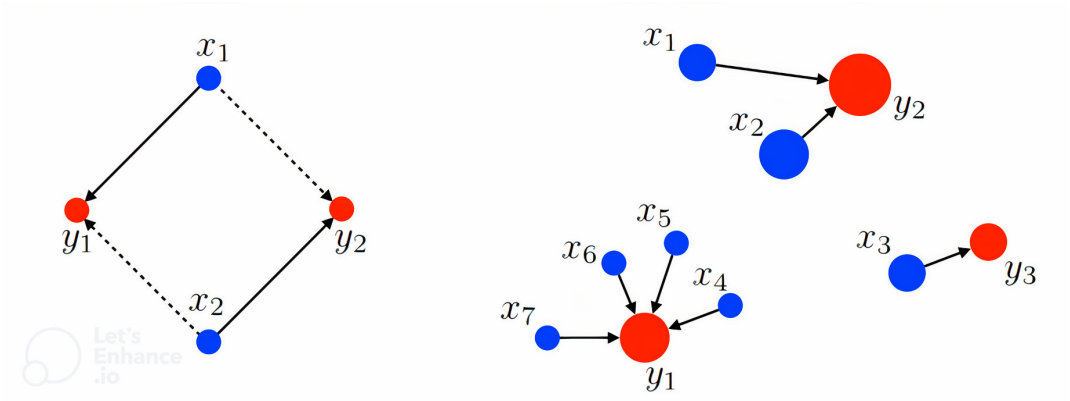


Figura 1.3: Izquierda: los puntos azules de la medida μ y los puntos rojos de la medida ν están equidistantes dos a dos. Por lo tanto, cualquiera de las asignaciones $\sigma = (1, 2)$ (línea completa) o $\sigma = (2, 1)$ (línea punteada) es óptima. Derecha: un *Monge map* puede asociar la medida azul μ a la medida roja ν. Los pesos α_i se muestran proporcionalmente al área del disco marcado en cada ubicación. El mapeo aquí es tal que $T(x_1) = T(x_2) = y_2$, $T(x_3) = y_3$, mientras que para $4 \leq i \leq 7$, se tiene que $T(x_i) = y_1$. (imagen obtenida de [4])

Observación 1.2.2. En el caso discreto, si $n = m$ y las medidas son una uniforme sobre su soporte, el problema de Monge corresponde a simplemente una permutación. Sin embargo, si $n \neq m$, entonces el problema de Monge puede **no tener solución**. Si se considera el caso en que $\mu = \delta_{x_1}$ y $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$, se tiene que $\nu(\{y_1\}) = \frac{1}{2}$, pero $\mu(T^{-1}(y_1)) \in \{0, 1\}$ dependiendo de cuándo $x_1 \in T^{-1}(y_1)$, por lo tanto, el transporte T no puede existir.

1.2.2. Formulación de Kantorovich

En la formulación de Monge las distribuciones se mueven según una función T, lo que impide dividir masa. En el caso dado en la observación 1.2.2 si se permitiera división de masa, se podría enviar la mitad de la masa de x_1 a y_1 y el resto a y_2. Esto es lo que hace la formulación de Kantorovich. Dado $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ se puede pensar que $d\pi(x, y)$ es la cantidad de masa transferida de x a y. De esta manera se puede mover masa desde x a multiples y.

Definición 1.5 (Problema de Kantorovich para vectores de probabilidad) *Dada una matriz de costos* $\mathbf{C} \in \mathbb{R}^{n \times m}$*, y dos vectores de probabilidad* $\mathbf{a} \in \Sigma_n$*,* $\mathbf{b} \in \Sigma_m$*, el problema de Kantorovich se formula de la siguiente forma:*

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle, \tag{1.8}$$

donde $\langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def}}{=} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}$*. Además,*

$$U(a, b) = \{ \quad P \in \mathbb{R}_+^{n \times m} \,|\, P1_m = a, P^\top 1_n = b\} \tag{1.9}$$

donde $1_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ *es un vector de unos, y se ha utilizado la siguiente notación de matriz-vector:*

$$P1_m = \left(\sum_j P_{i,j} \right)_i \in \mathbb{R}_+^n \qquad\qquad P^\top 1_n = \left(\sum_i P_{i,j} \right)_j \in \mathbb{R}_+^m$$

Definición 1.6 (Problema de Kantorovich para medidas discretas) *Dadas las medidas discretas* ($\mu = \sum_i \mathbf{a}_i \delta_{x_i}, \nu = \sum_j \mathbf{b}_j \delta_{y_j}$) *y una función de costos* $c(x, y)$*, se puede considerar la matriz de costos* $\mathbf{C}_{i,j} = c(x_i, y_j)$ *para definir el problema de Kantorovich sobre medidas discretas:*

$$\mathcal{L}_c(\mu, \nu) \stackrel{\text{def}}{=} \mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \tag{1.10}$$

Definición 1.7 (El problema de Kantorovich, caso general) *Dada una función de coste* $c(x, y) = \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *medible y acotada por abajo, dos medidas* μ, ν *se define el problema de transporte como*

$$\mathcal{L}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \tag{1.11}$$

Donde $\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \,|\, P_{\mathcal{X}\#}\pi = \mu, \; P_{\mathcal{Y}\#}\pi = \nu\}$ *donde* $P_{\mathcal{X}\#}$ *y* $P_{\mathcal{Y}\#}$ *son los push-forward de las proyecciones* $P_{\mathcal{X}}(x, y) = x$ *y* $P_{\mathcal{Y}}(x, y) = y$ *respectivamente. Se puede notar que la condición de la conservación de masa es equivalente a imponer que* $\pi(A \times \mathcal{Y}) = \mu(A)$ *y* $\pi(\mathcal{X} \times B) = \nu(B)$ *para todo conjunto medible* $A \subseteq \mathcal{X}$ *y* $B \subseteq \mathcal{Y}$*. Este conjunto corresponde a todos los* ***couplings*** *o* ***transport plans*** *entre* μ *y* ν*. Ilustraciones de diferentes* ***transport plans*** *se muestran en la Figura 1.4*

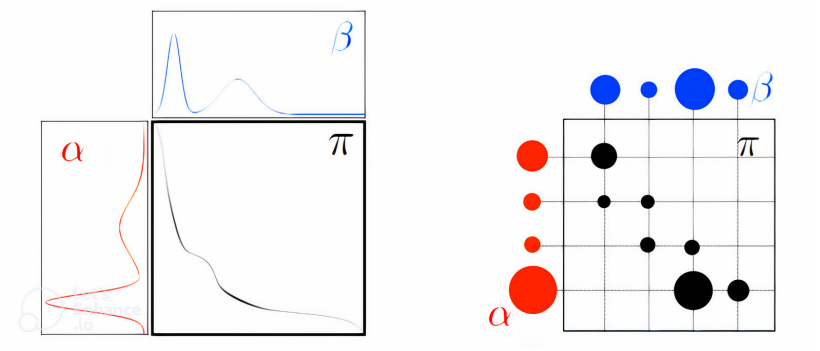


Figura 1.4: Izquierda: *coupling* continuo π que resuelve (1.11) entre dos medidas de una dimensión con densidad. El *coupling* está localizado a lo largo del gráfico del *Monge map* $(x, T(x))$ (mostrado en negro). Derecha: *coupling* discreto T que resuelve (1.6) entre dos medidas discretas. Las entradas positivas $T_{i,j}$ se muestran como discos negros en la posición (i, j) con radio proporcional a $T_{i,j}$. (imagen obtenida de [4])

Observación 1.2.3. El caso discreto es una situación especial, donde la medida producto está dada por $\pi = \sum_{i,j} \mathbf{P}_{i,j} \delta_{(x_i, y_j)}$.

Este problema podría no tener solución, sin embargo, el siguiente teorema garantiza, bajo ciertas hipótesis, cuando existe solución.

Teorema 1.2.4 (ver [8]). *(Existencia de un coupling óptimo). Sea $(\mathcal{X}, \mu)$ y $(\mathcal{Y}, \nu)$ dos espacios de probabilidad polacos; sea $a : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ y $b : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ dos funciones semicontinuas superiores tales que $a \in L^1(\mu)$ y $b \in L^1(\nu)$. Sea $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ una función de costo semicontinua inferior, tal que $c(x, y) \geq a(x) + b(y)$ para todo x, y. Entonces existe un coupling de (μ, ν) que minimiza el costo total* (1.11) *entre todos los coupling posibles.*

Observación 1.2.5. La hipotesis de que c está acotada por abajo garantiza que el valor del problema está bien definido en $\mathbb{R} \cup \{-\infty\}$. En la mayoría de los casos se considera $a = b = 0$

De manera general, en la teoría de la optimización, se puede definir lo que es un problema dual de un problema de optimización, que en esencia permite ver desde otro punto de vista el problema, y que suele resultar más fácil de resolver que el problema original. En el transporte óptimo existe la dualidad de Kantorovich, y en lo que sigue se asegura de que el problema dual tenga sentido, y más aún que relaciones tienen los valores de ambos problemas de optimización.

Teorema 1.2.6. *Para funciones de costo $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ semicontinuas inferior y acotadas inferiormente, se tiene*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\pi(x, y) = \sup_{f \in L^1(\mu),\, g \in L^1(\nu),\, f+g \leq c} \int_{\mathcal{X}} f(x) d\mu(x) + \int_{\mathcal{Y}} g(y) d\nu(y). \tag{1.12}$$

De manera equivalente se expresa como:

$$\mathcal{L}_c(\mu, \nu) = \sup_f \left(\int_{\mathcal{X}} f^c(x)\, d\mu(x) + \int_{\mathcal{Y}} f(y)\, d\nu(y) \right) \tag{1.13}$$

Con f una función continua acotada y $f^c(x) = \inf_{y \in \mathcal{Y}} \{c(x, y) - f(y)\}$

El siguiente Teorema proporciona una caracterización de los transportes para una función de costo particular.

Teorema 1.2.7 (Brenier, 1987)**.** *En el caso que $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ y $c(x,y) = \|x-y\|^2$, si al menos una de las 2 medidas involucradas tiene densidad con respecto a la medida de Lebesgue, entonces el óptimo π en el problema de Kantorovich es único y su soporte está en el grafo $\{(x, T(x))\ x \in \mathbb{R}^d\}$ de un Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$. Esto significa que $\pi = (id, T)_{\#}\mu$*

$$\forall h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}),\ \int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\,\mathrm{d}\pi(x,y) = \int_{\mathcal{X}} h(x, T(x))\,\mathrm{d}\mu(x) \tag{1.14}$$

. Además, la función T está únicamente definida como el gradiente de una función convexa ϕ, $T(x) = \nabla\phi(x)$ donde ϕ es la única, salvo constante, función convexa tal que $(\nabla\phi)_{\#}\mu = \nu$. Esta función está relacionada con el potencial dual f que resuelve 1.12 como $\phi(x) = \frac{\|x\|^2}{2} - f(x)$

Observación 1.2.8. El Teorema de Brenier asegura que en $\mathbb{R}^d$ y p=2 se tiene que para medidas con momento de orden 2 finito y al menos una de ellas tiene densidad, entonces los problemas de Monge y Kantorovich son equivalentes.

1.2.3. Distancia de Wasserstein

Supuesto 1.2.9. El espacio métrico $(\mathcal{X}, d)$ es un espacio geodésico localmente compacto, separable, dotado de una medida Borel σ-finita. En donde el concepto *geodésico* significa que el espacio $(\mathcal{X}, d)$ es completo y que cualquier par de puntos admite un punto medio con respecto a d.

Una característica importante de la Teoría de Transporte es que define una distancia entre histogramas y medidas de probabilidad tan pronto como la matriz de costos cumpla ciertas propiedades adecuadas. De hecho, OT puede entenderse como una forma canónica de elevar una distancia entre puntos a una distancia entre histogramas o medidas.

Definición 1.8 (Espacio de Wasserstein) *Dado un número $p \geq 1$, se define el p-espacio de Wasserstein $\mathcal{W}_p(\mathcal{X})$ por*

$$\mathcal{W}_p(\mathcal{X}) \stackrel{\text{def}}{=} \left\{\mu \in \mathcal{P}(\mathcal{X}) \mid \int_{\mathcal{X}} d(x, x_0)^p\,\mathrm{d}\mu(x) < +\infty,\ \textit{para algún } x_0 \in \mathcal{X}\right\} \tag{1.15}$$

donde $x_0 \in \mathcal{X}$ es arbitrario. Este espacio no depende de la elección de este punto.

El espacio de Wasserstein es el espacio de medidas de probabilidad que tienen *momento finito de orden p*

Definición 1.9 (Distancia de Wasserstein) *La p-distancia de Wasserstein entre μ y ν se define por medio de*

$$W_p(\mu,\nu) \stackrel{\text{def}}{=} \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p\,\mathrm{d}\pi(x,y)\right)^{\frac{1}{p}} = (\mathcal{L}_{d^p}(\mu,\nu))^{\frac{1}{p}} \tag{1.16}$$

Observación 1.2.10. Una aplicación de la desigualdad de Hölder prueba que

$$p \leq q \Rightarrow W_p \leq W_q \tag{1.17}$$

La distancia de Wasserstein W_p tiene muchas propiedades importantes, siendo la más relevante que es una distancia débil, es decir, permite comparar distribuciones singulares (por ejemplo, distribuciones discretas) cuyos soportes no se superponen y cuantificar el desplazamiento espacial entre los soportes de dos distribuciones.

1.2.4. Convergencia en el Sentido de Wasserstein

Dado que una distancia define una topología en un espacio métrico, a continuación se ven propiedades esenciales topológicas de este espacio.

Definición 1.10 ([8]) *[Convergencia débil en $\mathcal{P}_p$] Sea $(\mathcal{X}, d)$ un espacio métrico polaco, $p \in [1, \infty)$. Sea $(\mu_k)_{k\in\mathbb{N}}$ una sucesión de medidas de probabilidad en $\mathcal{P}_p(\mathcal{X})$ y sea μ otro elemento de este espacio. Entonces se dice que (μ_k) converge débil en $\mathcal{P}_p(\mathcal{X})$ si cualquiera de las siguientes propiedades equivalentes se satisfacen para algún (y por lo tanto para todo) $x_0 \in \mathcal{X}$:*

1. $\mu_k \to \mu$ *y* $\int d(x_0, x)^p d\mu_k \to \int d(x_0, x)^p d\mu$
2. $\mu_k \to \mu$ *y* $\limsup_{k\to\infty} \int d(x_0, x)^p d\mu_k \leq \int d(x_0, x)^p d\mu$
3. $\mu_k \to \mu$ *y* $\lim_{R\to\infty} \limsup_{k\to\infty} \int_{d(x_0,x)\geq R} d(x_0, x)^p d\mu_k = 0$
4. *para toda función continua ϕ con $|\phi(x)| \leq C(1 + d(x_0, x)^p), C \in \mathbb{R}$*

$$\int \phi(x) d\mu_k(x) \to \int \phi(x) d\mu(x)$$

Teorema 1.2.11 ([8])**.** *(W_p metriza $\mathcal{P}_p$) Sea $(\mathcal{X}, d)$ un espacio métrico polaco, y $p \in [1, \infty]$ entonces la distancia de Wasserstein W_p metriza la convergencia débil en $\mathcal{P}_p(\mathcal{X})$. En otras palabras, si $(\mu_k)_{k\in\mathbb{N}}$ es una sucesión de medidas en $\mathcal{P}(\mathcal{X})$ y μ otra medida en este mismo espacio, luego las proposiciones*

$$\mu_k \text{ converge débil a } \mu$$

y

$$W_p(\mu_k, \mu) \to 0$$

son equivalentes.

Observación 1.2.12. Como consecuencia del Teorema 1.2.11 la convergencia en el p-espacio de Wasserstein implica la convergencia de los momentos de orden p.

Corolario 1.2.13. *(Continuidad de W_p) Sea $(\mathcal{X}, d)$ un espacio métrico polaco, y $p \in [1, \infty]$ entonces la distancia de Wasserstein W_p es continua en $\mathcal{P}_p(\mathcal{X})$. Esto quiere decir que, si μ_k (resp ν_k) converge débil a μ en $\mathcal{P}_p(\mathcal{X})$ (análogo a ν) entonces*

$$W_p(\mu_k, \nu_k) \to W_p(\mu, \nu)$$

Observación 1.2.14. Si las convergencias anteriores son solo débiles, se puede concluir únicamente que $W_p(\mu, \nu) \leq \liminf W_p(\mu_k, \nu_k)$, esto quiere decir que la distancia de Wasserstein es semicontinua inferior en $\mathcal{P}(\mathcal{X})$

1.3. Motivación

La teoría de OT permite definir la noción de promedio de distribuciones de probabilidad, lo que se conoce como baricentro de Wasserstein (WB, por sus siglas en inglés)[1], el cual ha encontrado varias aplicaciones en *Machine Learning*, tales como interpolación de imágenes [9], *model ensembling* [10], entre otras. Por otra parte, el reciente desarrollo de la teoría de transporte óptimo débil (WOT, por sus siglas en inglés)[2] ha permitido formar la idea de baricentro débil de Wasserstein (WWB, por sus siglas en inglés) [3]. Este promedio posee diferentes propiedades matemáticas que son de especial interés. La principal propiedad es que el baricentro débil captura toda la información común que tienen las medidas de probabilidad involucradas, mientras que el baricentro usual se puede pensar como un promedio de puntos en dimensión finita.

Sin embargo, a diferencia de OT, la formulación débil WOT no presenta simetría. Vale decir, al resolver el problema, sí importa el orden en que se ocupan las medidas de probabilidad involucradas. Esta asimetría no ha recibido atención y es de principal interés conocer qué propiedades se preservan, cambian o surgen, ya que permite mayor interpretabilidad de este tipo de promedio, además de poder realizar comparaciones con los otros dos baricentros mencionados. Así como también dichas propiedades permiten, dado un problema, decidir qué tipo de promedio es útil según los requerimientos de la problemática que se busque resolver con estos métodos. En otras palabras, lo anterior se refiere a decidir qué tipo de promedio será conveniente usar dependiendo de qué información común es la que se busca capturar. Otro aspecto a abordar será el cálculo de estos baricentros, pues se sabe que estos problemas son difíciles de resolver numéricamente [11].

1.4. Hipótesis

Así como el baricentro débil tiene la propiedad de ser una variable latente común a todas las distribuciones promediadas, el baricentro débil reverso podría contener toda la información de las distribuciones involucradas, en otras palabras, cada distribución corresponde a una variable latente para el baricentro débil reverso. Esto puede ser útil en aplicaciones que requieran reunir toda la información de diferentes distribuciones de probabilidad en una sola, lo que proporcionaría ventajas por sobre los métodos existentes.

1.5. Objetivo General

Estudiar teórica y algorítmicamente los baricentros que se pueden formular a partir del transporte óptimo débil para aplicaciones en aprendizaje de máquinas, ventajas y desventajas de cada uno.

1.6. Objetivos Específicos

- Demostrar diferentes propiedades del baricentro débil reverso.
- Obtener algoritmo eficiente para el cálculo del baricentro débil y del reverso.
- Evaluar el comportamiento teórico de los tres promedios en dataset MNIST y gaussianas en una dimensión.

1.7. Estructura de la Tesis

Los capítulos restantes se organizan según lo explicado a continuación: el Capítulo 2 corresponde al Marco Teórico, en el que se define el transporte óptimo débil, junto a las respectivas formulaciones de baricentros. También se habla sobre Machine, Deep Learning, algunas arquitecturas de redes neuronales que se usaron y Algoritmos genéticos, necesarios para una mejor comprensión del trabajo realizado. En dicho capítulo también se exponen los principales algoritmos del estado del arte para resolver los problemas de optimización involucrados. En el Capítulo 3 se mostrarán los resultados teóricos y el diseño de los algoritmos para resolver los problemas de optimización considerados. En el Capítulo 4 se exhibirán los resultados numéricos, junto con un análisis de estos con asociaciones teóricas y algunas comparaciones entre los 3 baricentros. Finalmente, el Capítulo 5 muestra las conclusiones obtenidas una vez realizado el trabajo, en cuanto al cumplimiento de los objetivos y trabajo futuro.

Capítulo 2

Marco Teórico

En la literatura se señalan varias propiedades matemáticas relevantes relacionadas con el transporte óptimo. Asimismo, se ha desarrollado la teoría y algoritmos para poder computar lo que es el baricentro de Wasserstein. Dado lo anterior, en este capítulo se documenta la recopilación de los aspectos teóricos relevantes asociados al baricentro de Wasserstein y al área de transporte óptimo débil para poder desarrollar y comprender el baricentro débil reverso. Posteriormente, se mostrarán diferentes algoritmos que resuelven los problemas de optimización asociados.

2.1. Baricentros de Wasserstein

Por analogía con el caso euclidiano, donde el baricentro de los puntos $(x_1, \ldots, x_p)$ con coordenadas baricéntricas $(\lambda_1, \ldots, \lambda_p)$ se obtiene como el minimizador de $x \mapsto \sum_{i=1}^{p} \lambda_i \|x - x_i\|^2$, se propone el mismo procedimiento en el espacio de Wasserstein simplemente reemplazando la distancia euclidiana al cuadrado con la distancia al cuadrado de 2-Wasserstein. Formalmente, un baricentro corresponde a una *Frechet Mean* [12] en el espacio de Wasserstein. Las *Frechet Mean* generalizan la noción de promedio a espacios métricos generales.

Definición 2.1 (ver [1]) *Sea una familia finita de medidas* $\{\nu_i\}_{i=1,\ldots,n} \in \mathcal{P}_2(\mathbb{R}^d)$ *y pesos* $\lambda_1, \ldots, \lambda_n$ *tales que* $\sum_{i=1}^{n} \lambda_i = 1$ *el problema de baricentro se resume en encontrar una medida* μ *tal que*

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \nu_i) = \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} L(\mu) \tag{2.1}$$

La idea de los baricentros de Wasserstein es definir una interpolación no lineal entre varias medidas de probabilidad en $\mathbb{R}^d$.

2.1.1. Algoritmo de Punto Fijo

Se definirá una función $H : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathcal{P}_{2,ac}(\mathbb{R}^d)$ cuyos puntos fijos serán, bajo ciertas hipótesis, el baricentro de $\nu_1, \ldots, \nu_n \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. Sea $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ por el Teorema 1.2.7 existen los *Monge map* $T_1, \ldots, T_k$ tal que $T_j(X) \sim \nu_j$ donde $X \sim \mu$ y $W_2^2(\mu, \nu_j) = \mathbb{E}(\|X - T_j(X)\|^2)$ con todo lo anterior se define

$$H(\mu) := \left(\sum_{j=1}^{n} \lambda_j T_j \right) \# \mu \tag{2.2}$$

El cual corresponde a su vez un mapeo óptimo entre μ y $H(\mu)$ (una combinación positiva de mapeos óptimos es igual, según el Teorema 1.2.7 de Brenier, y la Observación 1.2.8, a la suma de gradientes de funciones convexas, que equivale al gradiente de una suma de funciones convexas, por lo tanto, óptimo nuevamente por el Teorema de Brenier).

Teorema 2.1.1 (ver [13]). *Si ν_j tiene densidad, entonces H mapea $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ en sí mismo y es continua con respecto a W_2*

Observación 2.1.2. En este contexto, se asumirá que $\nu_1, \ldots, \nu_n \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ al menos una de ellas tiene densidad acotada

Proposición 2.1.3 (ver [13]). *Sea $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ entonces*

$$L(\mu) \geq L(H(\mu)) + W_2^2(\mu, H(\mu)) \tag{2.3}$$

Corolario 2.1.4 (ver [13]). *Si μ es el único baricentro de $\nu_1, \ldots, \nu_n$ entonces $H(\mu) = \mu$*

Gracias a los resultados anteriores, se propone definir una sucesión de la siguiente manera

$$\mu_{k+1} = H(\mu_k) \tag{2.4}$$

Teorema 2.1.5 (ver [13]). *La sucesión definida en (2.4) es tensa. Toda subsucesión débilmente convergente también converge en W_2 a alguna medida de probabilidad en $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ la cual es un punto fijo de H. En particular, si H tiene un único punto fijo $\tilde{u}$, entonces $\tilde{u}$ es el baricentro de $\nu_1, \ldots, \nu_n$ y $W_2(\mu_n, \tilde{\mu}) \to 0$*

2.2. Transporte Óptimo Débil

Para pasar al problema de transporte óptimo débil, notar que cualquier *coupling* $\pi \in \Pi(\mu, \nu)$ se puede desintegrar de la siguiente manera

$$\pi(dxdy) = \mu(dx)\pi_x(dy)$$

Intuitivamente, π_x contiene toda la información de como la masa tomada de x se distribuye sobre el espacio $\mathcal{Y}$. Con esta desintegración de π se tiene que

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(dxdy) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} c(x,y)d\pi_x(dy) \right) d\mu(dx)$$

En el problema de Monge-Kantorovich, la masa que se transfiere de $\mathcal{X}$ a $\mathcal{Y}$ es penalizada solo a través de su costo medio $\int_{\mathcal{Y}} c(x,y)d\pi_x(dy)$, en cambio, el transporte óptimo débil permite considerar penalizaciones más generales sobre π_x. En la Figura 2.1 se puede observar que es lo que hace la distribución π_x. Antes de definir el problema, se verá qué tipos de funciones de costo son las que se usarán, lo que está dado por la siguiente definición.

Definición 2.2 *Decimos que $C : \mathcal{X} \times \mathcal{P}_p(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ satisface la propiedad (A) si se cumplen las siguientes condiciones:*

1. *C es semicontinua inferiormente con respecto a la topología producto en $\mathcal{X} \times \mathcal{P}_p(\mathcal{Y})$.*
2. *C está acotada inferiormente.*

Si además la función $p \mapsto C(x,p)$ es convexa, es decir, para todo $x \in \mathcal{X}$ y $p,q \in \mathcal{P}_p(\mathcal{Y})$, se tiene que $C(x, \lambda p + (1-\lambda)q) \leq \lambda C(x,p) + (1-\lambda)C(x,q)$, para todo $\lambda \in [0,1]$. Se dice que la función C satisface la condición A^+.

Formalmente, se define el problema de transporte óptimo débil de la siguiente manera:

Definición 2.3 *Sea $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ una función de coste. $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ el problema de transporte óptimo débil está dado por*

$$V_c(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} c(x, \pi_x) d\mu(x) \tag{2.5}$$

.

Lo primero es verificar que el problema tenga sentido, y este resultado lo dan los Teoremas a continuación.

Teorema 2.2.1 (ver [14] Teorema 3.2 (Existencia y semicontinuidad))**.** *El ínfimo en WOT se alcanza y el valor $V_C(\mu,\nu)$ depende de manera semicontinua inferior de los marginales $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}_p(\mathcal{Y})$.*

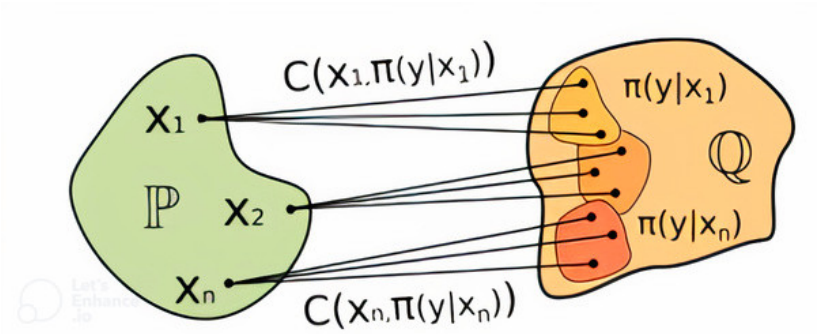


Figura 2.1: Ejemplo de WOT (imagen obtenida de [15])

Teorema 2.2.2 (ver [16] Teorema 1.1 (Existencia I))**.** *Sea $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ semicontinua inferior en ambas variables, acotada inferiormente y convexa en el segundo argumento. Entonces, el problema*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} c(x, \pi_x) d\mu(x) \tag{2.6}$$

admite un minimizador.

Teorema 2.2.3 (ver [16] Teorema 1.2 (Existencia II))**.** *Sea $\nu \in \mathcal{P}_p(\mathcal{Y})$. Sea $C : \mathcal{X} \times \mathcal{P}_d(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ semicontinua inferior con respecto a la topología producto en $\mathcal{X} \times \mathcal{P}_d(\mathcal{Y})$, acotada inferiormente y convexa en el segundo argumento. Entonces, el problema*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} c(x, \pi_x) d\mu(x) \tag{2.7}$$

admite un minimizador.

Observación 2.2.4. El Teorema 2.2.2 es un caso especial del Teorema 2.2.3. Para ver esto, basta considerar d_Y como una métrica acotada compatible con la topología original de $\mathcal{Y}$.

También se tiene que si c es estrictamente convexa en el segundo argumento y $V_c(\mu,\nu) < \infty$, entonces el minimizador $\pi^* \in \Pi(\mu,\nu)$ es único. Aquí nace una primera gran diferencia entre esta formulación y la del transporte óptimo, y es que para asegurar la existencia de solución no se requiere imponer condiciones sobre las medidas de probabilidad involucradas.

La siguiente definición es de gran utilidad y se usará bastante a lo largo de este trabajo:

Definición 2.4 (ver [17]) *(Orden convexo de medidas) dos medidas están en orden convexo* $\mu \leq_c \nu$ *si*

$$\int \phi \mathrm{d}\mu \leq \int \phi \mathrm{d}\nu$$

para toda función ϕ *convexa.*

Un caso particular de orden convexo entre medidas relevante para este trabajo está dado en la observación 2.2.5

Observación 2.2.5. En el caso de distribuciones gaussianas unidimensionales, se tiene que $\mu = \mathcal{N}(\alpha, \sigma_1^2) \leq_c \nu = \mathcal{N}(\beta, \sigma_2^2)$ sí y solo sí $\alpha = \beta$ y $\sigma_1 \leq \sigma_2$

Teorema 2.2.6 (Strassen [17])**.** *Sean* $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^n)$*; las siguientes proposiciones son equivalentes:*

1. $\mu \leq_c \nu$,
2. *Existe una martingala* (X_0, X_1) *tal que* $X_0 \sim \mu$ *y* $X_1 \sim \nu$.

Este teorema otorga una caracterización útil para verificar el orden convexo entre medidas.

La función de coste a considerar es la baricéntrica, esto es

$$c_0(x,p) = \|x - \int y dp(y)\|^2 \tag{2.8}$$

con ello se define el funcional

$$V(\mu|\nu) = V_{c_0}(\mu,\nu) \tag{2.9}$$

y es de principal interés por el siguiente teorema:

Teorema 2.2.7 (ver [18])**.** *Sea* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ *y* $\nu \in \mathcal{P}_1(\mathbb{R}^d)$*. Entonces existe un único* $\eta^* \leq_c \nu$ *tal que*

$$W_2^2(\mu,\eta^*) = \inf_{\eta \leq_c \nu} W_2^2(\mu,\eta) = V(\mu|\nu) \tag{2.10}$$

Más aún, existe una función convexa $\psi : \mathbb{R}^d \to \mathbb{R}$ *de clase* C^1 *con* $\nabla\psi$ *1-Lipschitz tal que* $\nabla\psi_{\#} = \eta^*$*. Finalmente, el coupling óptimo* $\pi^{\mu,\nu} \in \Pi(\mu,\nu)$ *verifica* $\int y d\pi_x^{\mu,\nu}(y) = \nabla\psi(x) \quad \mu$ *c.s*

Teorema 2.2.8 (ver [18], Teorema 1.5)**.** *Sea* $1 \leq \rho < \infty$*,* $(\mu_k)_{k\in\mathbb{N}} \in \mathcal{P}_p(\mathbb{R}^d)$*,* $(\nu_k)_{k\in\mathbb{N}} \in \mathcal{P}_1(\mathbb{R}^d)$*. Si* $\mu_k \to \mu$ *en* W_ρ *y* $\nu_k \to \nu$ *en* W_1*, entonces* $\lim_k V_(\mu_k|\nu_k) = V(\mu|\nu)$*.*

2.3. Formulación Dual WOT

Así como existe la formulación dual para el problema de Kantorovich vista en el Teorema 1.2.6, también se puede formular la versión dual para el transporte óptimo débil. Para ello,

primero se darán definiciones pertinentes con el fin de formular el problema dual.

Se denota por Φ_t al conjunto de funciones continuas en $\mathcal{Y}$ que satisfacen la restricción de crecimiento:

$$\exists y_0 \in \mathcal{Y}, \exists a, b \in \mathbb{R}^+, \forall y \in \mathcal{Y} : |\psi(y)| \leq a + b d_{\mathcal{Y}}(y, y_0)^t$$

y por $\Phi_{b,t}$ al subconjunto de funciones en Φ_t que están acotadas inferiormente. Además, se define la noción de C-conjugada:

Definición 2.5 *La C-conjugada de una función medible $f : \mathcal{Y} \to \mathbb{R}$, denotada por $R_C f$, está dada por*

$$R_C f(x) := \inf_{p \in \mathcal{P}_p(\mathcal{Y})} \int_{\mathcal{Y}} f(y) dp(y) + C(x, p) \tag{2.11}$$

Observación 2.3.1. Nótese que para costos C fuertes, el ínfimo se alcanza para cualquier $\mu \in \mathcal{P}(\mathcal{Y})$ con soporte en el conjunto $\{\arg\min_{y \in \mathcal{Y}}(c(x, y) - f(y))\}$. Por lo tanto, basta con utilizar la transformada c fuerte:

$$R_C f(x) = f^c(x) = \inf_{y \in \mathcal{Y}} \{c(x, y) - f(y)\}.$$

La dualidad en el caso débil toma la siguiente forma, que se asemeja (y generaliza) a (1.12).

Teorema 2.3.2. *(Dualidad de Kantorovich para transporte débil)[ver [16], Teorema 3.1]. El problema de transporte débil WOT admite la representación dual*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} C(x, \pi_x)\, d\mu(x) = \sup \int_{\mathcal{X}} f(x) d\mu(x) + \int_{\mathcal{X}} g(y) d\nu(y) \tag{2.12}$$

donde el supremo se toma sobre funciones $f \in L^1(\mu)$, $g \in \Phi_t(\mathcal{Y})$, que satisfacen $f(x) + \int_{\mathcal{Y}} g(y) dp(y) \leq C(x, p)$ para todo $x \in \mathcal{X}$, $p \in \mathcal{P}_p(\mathcal{Y})$.

Más aún, la dualidad se puede reescribir en función de la C-conjugada, tal como se muestra a continuación.

Teorema 2.3.3 (ver [16])**.** *Sea $C : \mathcal{X} \times \mathcal{P}_p(\mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ que satisface la Condición (A+). Entonces,*

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}} C(x, \pi_x)\, d\mu(x) = \sup_{\psi \in \Phi_{b,t}} -\int_{\mathcal{Y}} \psi(y) d\nu(y) + \int_{\mathcal{X}} R_C \psi(x)\, d\mu(x) \tag{2.13}$$

Observación 2.3.4. Notar que la optimización en (2.13) es equivalente a optimizar sobre funciones $-\psi \in \Phi_{b,t}$ y reemplazar ψ por $-\psi$ en los lugares que corresponda.

2.3.1. Reformulación del Problema Dual

Primero, se reescribe la optimización en la transformación (2.11). Para esto, sea $Z \subset \mathbb{R}^d$ con una distribución no atómica S sobre él. Por ejemplo, $S = \mathcal{N}(0, I)$. En lo que sigue, todos los resultados sobre la reformulación del problema dual usan lo mencionado en la Observación 2.3.4

Lema 2.3.5 (ver[15]).

$$R_C f(x) = \inf_t C(x, t_\# S) - \int_Z f(t(z)) dS(z) \tag{2.14}$$

donde el ínfimo se toma sobre todas las funciones $t : Z \to \mathcal{Y}$ medibles

Lema 2.3.6 (ver [15]).

$$\int_{\mathcal{X}} R_C f(x)\, d\mu(x) = \inf_T \int_{\mathcal{X}} \left(C(t, T(x, \cdot)_\# S) - \int_Z f(T(x, z)) dS(z) \right) d\mu(x) \tag{2.15}$$

donde el ínfimo se toma sobre todas las funciones $T : \mathcal{X} \times Z \to \mathcal{Y}$ medibles

Corolario 2.3.7 (ver [15]). *(Reformulación minimax del problema dual).*

$$V_C(\mu, \nu) = \sup_f \inf_T \mathcal{L}(f, T) \tag{2.16}$$

donde la función $\mathcal{L}$ está definida por

$$\mathcal{L}(f, T) = \int_{\mathcal{Y}} f(y)\, d\nu(y) + \inf_T \int_{\mathcal{X}} \left(C(t, T(x, \cdot)_\# S) - \int_Z f(T(x, z)) dS(z) \right) d\mu(x) \tag{2.17}$$

Se dice que las funciones $T : \mathcal{X} \times Z \to \mathcal{Y}$ son mapas estocásticos. Si un mapa T es independiente de z, es decir, para todo $(x, z) \in \mathcal{X} \times Z$ se tiene que $T(x, z) \equiv T(x)$, decimos que el mapa es determinista. Un *transport plan* π^* podría ser no determinista, es decir, podría no existir una función determinista $T : \mathcal{X} \to \mathcal{Y}$ que satisfaga $\pi^* = [\mathrm{id}_X, T]\#\mu$. Sin embargo, cada *transport plan* $\pi \in \Pi(\mu, \nu)$ puede ser representado implícitamente mediante una función estocástica $T : \mathcal{X} \times Z \to \mathcal{Y}$. Este hecho se conoce como *noise outsourcing* ([19] Teorema 5.10) para $Z = [0, 1] \subset \mathbb{R}^1$ y $S = \text{Uniforme}([0, 1])$.

Con base en la formulación anterior, en [15] se propone usar redes neuronales T_θ, f_ω para parametrizar T y f respectivamente. Para encontrar los parámetros óptimos se usa un algoritmo llamado *stochastic gradient ascent-descent (SGAD)* cuyo objetivo es resolver el problema minimax propuesto en 2.3.7. Los detalles se encuntran en 1. Cabe destacar que este algoritmo se puede aplicar tanto para el caso débil como el que no, y que la función $\mathcal{L}$ que aparece en el algoritmo corresponde a la dada en la ecuación (2.17).

Algorithm 1 Transporte Óptimo Neural (NOT)

Require: Distribuciones P, Q, S accesibles mediante muestras, red de mapeo $T_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, red potencial $f_\omega : \mathbb{R}^d \to \mathbb{R}$, número de iteraciones internas K_T, costo (débil) $C : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$, estimador empírico $\hat{C}(x, Z)$ para el costo.

Ensure: Mapeo estocástico de OT aprendido T_θ que representa un plan de OT entre las distribuciones P, Q.

repeat
  Muestrear lotes $Y \sim Q$, $X \sim P$
  for cada $x \in X$ **do**
    Muestrear lote $Z_x \sim S$
    $\mathcal{L}_f \leftarrow \frac{1}{|X|} \sum_{x \in X} \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega(T_\theta(x, z)) - \frac{1}{|Y|} \sum_{y \in Y} f_\omega(y)$
    Actualizar ω usando $\frac{\partial \mathcal{L}_f}{\partial \omega}$;
  end for
  for $k_T = 1, 2, \ldots, K_T$ **do**
    Muestrear batches $X \sim \mathbb{P}$
    Para cada $x \in X$ muestrear batch $Z_x \sim \mathbb{S}$
    $\mathcal{L}_T \leftarrow \frac{1}{|X|} \sum_{x \in X} \hat{C}(x, T_\theta(x, z)) - \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega(T_\theta(x, z))$
    Actualizar θ usando $\frac{\partial \mathcal{L}_T}{\partial \theta}$
  end for
until alguna condición

2.4. Baricentros Débiles de Wasserstein

De manera análoga al baricentro de Wasserstein, se seguirá la misma idea para definir el baricentro débil de Wasserstein.

Definición 2.6 (ver [3]) *Sea una familia finita de medidas* $\{\nu_i\}_{i=1,\ldots,n} \in \mathcal{P}_2(\mathbb{R}^d)$ *y pesos* $\lambda_1, \ldots, \lambda_n$ *tales que* $\sum_{i=1}^n \lambda_i = 1$ *el problema de baricentro se resume en econtrar una medida* μ *tal que*

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i V(\mu | \nu_i) = \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} N(\mu) \tag{2.18}$$

En otras palabras, (2.18) el baricentro débil promedia, con respecto a la distancia de Wasserstein, una elección óptima de medidas de probabilidad $\{\eta_1, \ldots, \eta_n\}$ que están más concentradas que las ν_i en el sentido que $\eta_i \leq_c \nu_i \ \forall i$.

Proposición 2.4.1 (ver [3]). *El problema de baricentro débil en* (2.18) *admite un minimizador* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$

En lo que sigue, se denotará por X e Y_i a las variables aleatorias con leyes μ y ν_i respectivamente.

Lema 2.4.2 (ver [3]). *Si* μ *es un baricentro débil de* $\{\nu_i\}_{i=1,\ldots,n}$ *y* $\mu' \leq_c \mu$ *entonces* $\mu\prime$ *también es un baricentro débil. En particular, la masa de Dirac con soporte en* $\mathbb{E}_\mu(X)$ *siempre es un baricentro débil. Más aún, la distribución de Dirac* δ_ω *es un baricentro débil si y solo si* $\omega = \sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i}(Y_i)$

Una consecuencia del lema anterior es que para cualquier baricentro débil μ

$$\mathbb{E}_\mu(X) = \sum_{i=1}^{n} \lambda_i \mathbb{E}_{\nu_i}(Y_i) \tag{2.19}$$

Y el valor del problema (2.18) está dado por

$$\inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i V(\mu|\nu_i) = \sum_{i=1}^{n} \lambda_i \|\mathbb{E}_{\nu_i}(Y_i)\|^2 - \| \sum_{i=1}^{n} \lambda_i \mathbb{E}_{\nu_i}(Y_i)\|^2 \tag{2.20}$$

Proposición 2.4.3 (ver [3]). *Una medida $\mu \in \mathcal{P}(\mathbb{R}^d)$ es un baricentro débil de $\{\nu_i\}_{i=1,\dots,n}$ si y solo si su media satisface* (2.19) *y $\tilde{\mu} \leq_c \tilde{\nu}_i$ para todo $1 \leq i \leq n$ donde $\tilde{\nu}$ denota la versión centrada de la ley ν*

Lema 2.4.4 (ver [3]). *Sean $\nu_1, \dots, \nu_n \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ al menos una de ellas con densidad acotada. $\tilde{\mu}$ y $\hat{\mu}$ denotan el débil y el baricentro de Wasserstein. Entonces*

$$W_2^2(\tilde{\mu}, \hat{\mu}) \leq 2 \sum_{i=1}^{n} \lambda_i (\mathbb{E}\|Y_i\|^2 - \|\mathbb{E}Y_i\|^2) \tag{2.21}$$

Una definición importante para el resto del trabajo es la siguiente

Definición 2.7 *La proyección baricéntrica se define como*

$$S_\mu^\nu : x \to \int y \, \mathrm{d}\pi_x^{\mu,\nu}(y) \tag{2.22}$$

para $\pi^{\mu,\nu} \in \Pi(\mu, \nu)$ alcanzado el mínimo en el WOT.

Una propiedad bien importante es que el baricentro débil codifica toda la información común geométrica presente en todas las medidas a promediar consideradas, por lo tanto, se puede interpretar como la distribución de una variable latente subyacente a las realizaciones de variables aleatorias con leyes ν_i para todo $1 \leq i \leq n$. Esto se formaliza en el siguiente teorema:

Teorema 2.4.5 (ver [3]). *Sea μ un baricentro débil de $\{\nu_i\}_{i=1,\dots,n}$. Entonces para cada $1 \leq i \leq n$, las variables aleatorias $Y_i \sim \nu_i$ pueden ser realizadas como*

$$Y_i = X + (\mathbb{E}(Y_i) - \mathbb{E}(X)) + \tilde{Y}_i \tag{2.23}$$

donde $X \sim \mu$ y $\tilde{Y}_i = Y_i - \mathbb{E}(Y_i|X)$ es centrada condicionada en X. Más aún, se tiene que $S_\mu^\nu(X) = X + (\mathbb{E}(Y_i) - \mathbb{E}(X))$ para todo $1 \leq i \leq n$. Finalmente, $\mathbb{E}(Y_i - \mathbb{E}(Y_i)|X - \mathbb{E}(X)) = X - \mathbb{E}(X)$, o de manera equivalente, $\hat{\mu} \leq_c \hat{\nu}_i$ con $\hat{\mu}$ y $\hat{\nu}$ las leyes de $X - \mathbb{E}(X)$ e $Y_i - \mathbb{E}(Y_i)$ respectivamente.

El teorema anterior quiere decir que cada $Y_i \sim \nu_i$ se puede realizar sampleando una variable aleatoria X común a todas, distribuyendo según el baricentro débil μ, trasladando este valor por $\mathbb{E}(Y_i) - \mathbb{E}(X)$ y añadiendo una *componente específica* $\tilde{Y}_i$ o ruido idiosincrático, centrado y condicionado en X.

2.4.1. Algoritmo de Punto Fijo

Como se vio en la sección 2.1.1 para el problema de baricentro de Wasserstein existe un algoritmo iterativo basado en los puntos fijos de la función L. Para el baricentro débil se mostrará una metodología similar, pero usando las proyecciones baricéntricas en el problema de transporte óptimo débil, lo cual es válido para cualquier distribución.

Dadas $\nu_1, \dots, \nu_n \in \mathcal{P}_2(\mathbb{R}^d)$ se define

$$\mu_{k+1} = G(\mu_k), \text{ con } G(\mu) := \left(\sum_{j=1}^{n} \lambda_j S_\mu^\nu \right) \# \mu \tag{2.24}$$

Una diferencia fundamental entre el algoritmo de punto fijo para el baricentro de Wasserstein y el débil es el hecho que en el problema de transporte se verifica que $T_\# \mu = \nu$ mientras que la medida $S_\mu^\nu \# \mu$ sigue dependiendo de μ.

Teorema 2.4.6 (ver [3]). *La función H es continua con respecto a W_2 de $\mathcal{P}_2(\mathbb{R}^d)$ en $\mathcal{P}_2(\mathbb{R}^d)$*

Corolario 2.4.7 (ver [3]). *Si μ es un baricentro débil de $\nu_1, \dots, \nu_n$ entonces $G(\mu) = \mu$, es decir, $x = \sum_{i=1}^n \lambda_i \mathcal{S}_\mu^{\nu_i}(x), \mu(x) ctp$*

Proposición 2.4.8 (ver [3]). *Sea $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ entonces*

$$N(\mu) \geq N(G(\mu)) + W_2^2(\mu, G(\mu)) \tag{2.25}$$

Teorema 2.4.9 (ver [3]). *La sucesión definida en (2.24) y partiendo desde $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ es tensa y toda subsucesión débilmente convergente también converge en W_2 a algún punto fijo de G.*

2.5. Machine Learning

El Aprendizaje de Máquinas o *Machine Learning* (ML) es una disciplina de la Inteligencia Artificial que engloba a distintos algoritmos que permiten que un sistema aprenda a realizar una tarea de forma autónoma o sin ser explícitamente programada para aquello. Típicamente, las tareas aprendidas mediante algoritmos de ML son las de clasificación, regresión, entre otras. Dentro de la disciplina existen distintos enfoques de aprendizaje correspondientes al aprendizaje supervisado no supervisado, semi-supervisado y reforzado.

2.6. Deep Learning

El Aprendizaje Profundo o *Deep Learning* (DL) corresponde a un subconjunto de algoritmos de ML que se basa en redes neuronales de varias capas. En los últimos años, se han desarrollado distintas variantes de algoritmos de ML que se ajustan a las tareas específicas que se busca resolver con esta clase de algoritmos. A continuación, se entrega una explicación de los algoritmos que son mencionados y utilizados en este trabajo:

2.6.1. El Perceptrón

El perceptrón es un modelo probabilístico propuesto por Frank Rosenblatt [20]. Este modelo corresponde a la versión más básica de una red neuronal y matemáticamente se define

de la siguiente manera:

Definición 2.8 *(Perceptrón) El perceptrón es una función $f : \mathbb{R}^d \to \mathbb{R}$ tal que, para $x, \omega, \in \mathbb{R}^d$ y $b \in \mathbb{R}$*

$$f(x) = \begin{cases} 1 & si \quad w^T x + b \geq 0 \\ \\ 0 & si \quad w^T x + b < 0 \end{cases}$$

2.6.2. Perceptrón Multicapa

Definición 2.9 *(Perceptrón multicapa) Este tipo de red neuronal consiste en una función $f : \mathbb{R}^d \to \mathbb{R}^q$ definida recursivamente de la siguiente manera*

$$h_0 = x \tag{2.26}$$
$$h_i = g_i(W_i h_{i-1} + b_i) \tag{2.27}$$
$$h_M = W_M h_{M-1} + b_M \tag{2.28}$$
$$f(x) = h_M \tag{2.29}$$

Considerando $M > 1$, $K_0 = d$, $K_i \in \mathbb{N}$ para $i \in \{1, \ldots, M-1\}$ y $K_M = q$ se tiene que $W_i \in \mathbb{R}^{K_i \times \mathcal{K}_{i-1}}$, $b_i \in \mathbb{R}^{K_i}$ h_0 se le conoce como la capa de entrada, para cada $i \in \{1, \ldots, M-1\}$ h_i corresponde a las capas ocultas y h_M la capa de salida. Se abrevia MLP. Por último, para cada $i \in \{1, \ldots, M-1\}$ g_i es una función de activación, cuyo objetivo es hacer que las redes neuronales sean funciones no lineales.

Observación 2.6.1. La función de activación más popular debido a su buen rendimiento es la función $ReLU : \mathbb{R} \to \mathbb{R}$, para $x \in \mathbb{R}$:

$$ReLU(x) = max(0, x).$$

2.7. Redes Neuronales Convolucionales

Las *Convolutional Neural Networks* (CNN) [21] o Redes Neuronales Convolucionales son un tipo de red neuronal que fueron propuestas inicialmente para el procesamiento de imágenes. Basan su funcionamiento en el aprendizaje de filtros o kernels que son aplicados sobre la imagen mediante la operación de convolución. Esta operación se define como un producto escalar que se desplaza sobre la imagen, obteniendo un valor para cada una de las posiciones posibles. Los componentes clave de una CNN incluyen capas convolucionales, capas de agrupación y capas totalmente conectadas. Las cuales se detallan a continuación:

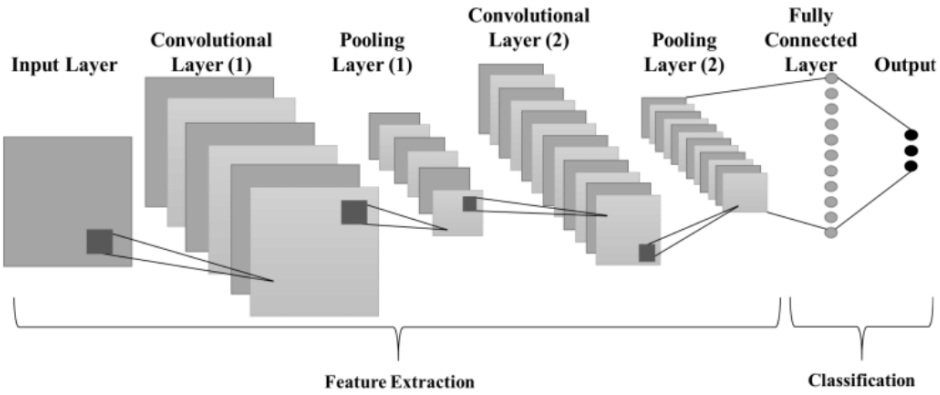


Figura 2.2: Arquitectura general de una CNN (imagen obtenida de [22])

- Capas Convolucionales: Estas capas aplican operaciones de convolución a los datos de entrada. La convolución implica deslizar una pequeña ventana (llamada kernel o filtro) sobre la entrada para realizar multiplicaciones y sumas elemento por elemento, capturando patrones locales.
- Capas de Agrupación o *Pooling*: Las capas de agrupación reducen las dimensiones espaciales de los datos, disminuyendo la complejidad computacional y capturando las características más relevantes. Las operaciones de agrupación comunes incluyen agrupación máxima y agrupación promedio.
- Capas Totalmente Conectadas o *Fully Connected*: Estas capas conectan cada neurona con todas las neuronas en las capas anteriores y siguientes, permitiendo que la red haga predicciones basadas en las características aprendidas.

En general cada una de estas capas se entrelazan entre sí tal como se ilustra en la figura 2.2. La disposición jerárquica de estas capas permite que las CNN aprendan automáticamente representaciones jerárquicas de características, desde bordes y texturas simples hasta estructuras complejas en las imágenes.

2.7.1. ResNet

Cuando se añaden más capas, surge un problema típico del aprendizaje profundo conocido como el gradiente que desaparece/explota. Esto provoca que el gradiente se vuelva cero o excesivamente grande. Por lo tanto, la tasa de error de entrenamiento y prueba aumenta de manera similar a medida que se incrementa el número de capas, tal como se muestra en la Figura 2.3

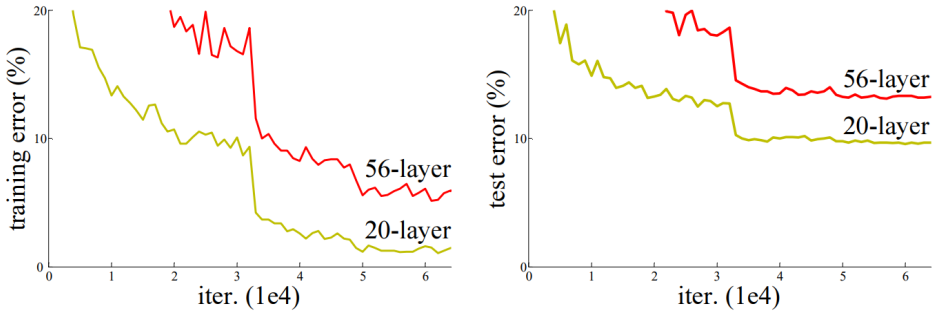


Figura 2.3: Error de entrenamiento (izquierda) y error de prueba (derecha) en CIFAR-10 con redes CNN de 20 capas y 56 capas. La red más profunda tiene un error de entrenamiento más alto y, por lo tanto, un error de prueba más alto. (imagen obtenida de [23])

ResNet, que significa Red Neuronal Residual [23], fue introducida para abordar el desafío de entrenar redes neuronales muy profundas y el problema del gradiente que desaparece/explota. Una de las innovaciones clave en ResNet es el uso de bloques residuales, que incluyen conexiones de salto (o conexiones directas). La conexión de salto evita algunos niveles intermedios para vincular las activaciones de capas con las capas subsecuentes. Esto crea un bloque residual. Estos bloques residuales se apilan para crear ResNets.

La estrategia detrás de esta red es permitir que la red ajuste el mapeo residual en lugar de hacer que las capas aprendan el mapeo subyacente. Por lo tanto, deja que la red ajuste en lugar de utilizar. Por ejemplo, el mapeo inicial de $H(x)$ se emplea $H(x) = F(x) + x$, tal como se muestra en la Figura 2.4

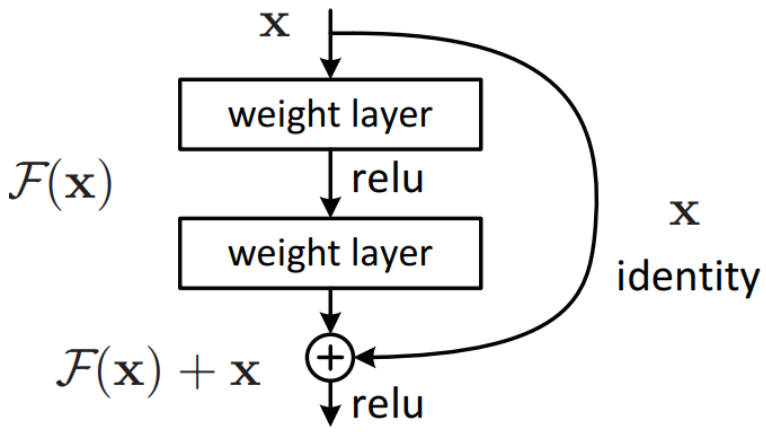


Figura 2.4: conexión de salto. (imagen obtenida de [23])

El beneficio de incluir este tipo de enlace de salto es que la regularización omitirá cualquier capa que degrade el rendimiento de la arquitectura. Como resultado, es posible entrenar una red neuronal extremadamente profunda sin encontrarse con problemas de gradientes que desaparecen o se expanden.

2.7.2. U-Net

U-Net es una arquitectura de red neuronal convolucional diseñada para tareas de segmentación semántica, donde el objetivo es clasificar cada píxel en una imagen en diferentes clases. La característica distintiva de U-Net es su estructura en forma de U, que incluye una ruta de contracción (codificador) y una ruta de expansión (decodificador). El codificador captura información contextual, mientras que el decodificador recupera información espacial. Los componentes clave de la arquitectura U-Net incluyen:

- Ruta de Contracción (Codificador): El codificador consiste en una serie de capas convolucionales y de agrupación que reducen las dimensiones espaciales de la imagen de entrada, capturando características de alto nivel.
- Cuello de Botella: El cuello de botella es una capa central que retiene una representación comprimida de las características de entrada.
- Ruta de Expansión (Decodificador): El decodificador consta de capas de aumento de muestreo y convolucionales que restauran las dimensiones espaciales. Se introducen conexiones de salto para concatenar mapas de características desde el codificador hasta el decodificador, facilitando la recuperación de detalles finos.
- Capa Final: La capa final generalmente emplea una capa convolucional con una función de activación softmax para la clasificación píxel a píxel.

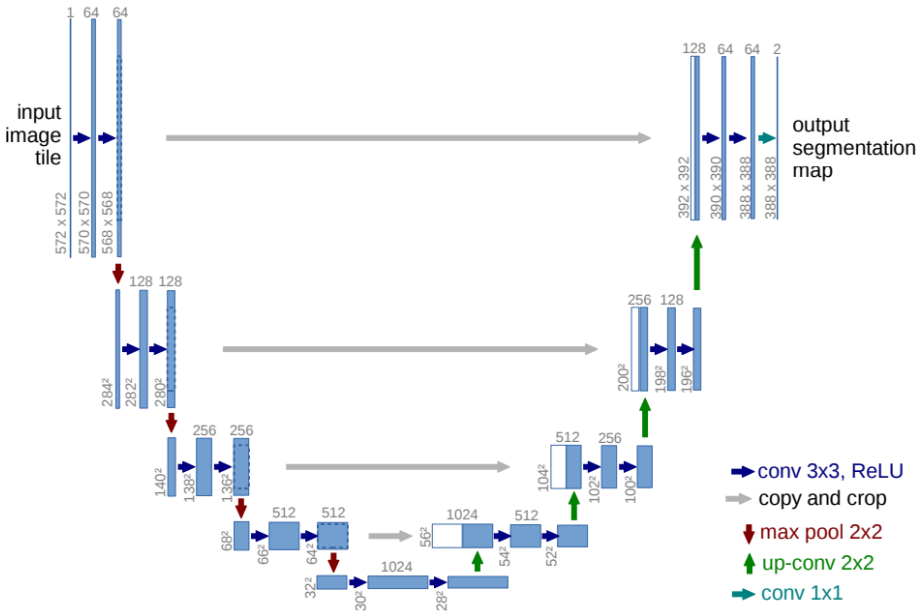


Figura 2.5: Arquitectura U-net (ejemplo para 32x32 píxeles en la resolución más baja). Cada caja azul corresponde a un mapa de características multicanal. El número de canales se indica en la parte superior de la caja. El tamaño x-y se proporciona en la esquina inferior izquierda de la caja. Las cajas blancas representan mapas de características copiados. Las flechas indican las diferentes operaciones. (imagen obtenida de [24])

Esta formulación representa la aplicación secuencial de los componentes del codificador,

el cuello de botella y el decodificador en la arquitectura U-Net, tal como se muestra en la Figura 2.5.

2.8. Algoritmos Genéticos

Dado que para calcular baricentros se debe resolver un problema de optimización, se tomará el enfoque de emplear algoritmos genéticos para resolverlos. Su elección radica en que se basan solo en evaluaciones de la función objetivo para encontrar soluciones, y son lo suficientemente flexibles para poder incorporar las restricciones de los problemas involucrados. Esta sección está enfocada en explicar en qué consisten. Los algoritmos genéticos son algoritmos estocásticos para resolver problemas de optimización basados en la teoría de la evolución [25]. Estos cuentan con una población inicial, donde cada individuo representa una posible solución del problema, y mediante operaciones genéticas hacen evolucionar la población inicial de individuos sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica (mutaciones y recombinaciones genéticas). Además, se realiza una selección de acuerdo con algún criterio, en función del cual se decide qué individuos son los más adaptados y sobreviven, y cuáles son menos aptos, que van siendo descartados. Mediante este proceso se van obteniendo nuevas soluciones, y basándose en la idea de que los más aptos son los que sobreviven, y en un problema de optimización apto quiere decir cercano al óptimo, entonces al terminar el procedimiento se estaría cerca o en el óptimo.

Cada individuo debe estar codificado usando los caracteres de algún alfabeto, y estas codificaciones representan una solución del problema. En la construcción de estos algoritmos se pueden presentar diversas variaciones, dependiendo de cómo se decide el reemplazo de los individuos para formar la nueva población. En general, el pseudocódigo consiste de los siguientes pasos:

1. **Inicialización:** Se genera aleatoriamente la población inicial Q_0, que está constituida por individuos que representan soluciones del problema.

2. **Evaluación:** A cada uno de los individuos de esta población se aplicará la función de aptitud para saber que tan buena es la solución que se está codificando.

3. **Selección:** Se seleccionan aleatoriamente M individuos de la población, aceptando repeticiones.La probabilidad de selección $p_s : Q_t \to [0, 1]$ es proporcional a su adaptabilidad. Esta es:
$$p_s(x) = \frac{f_{\text{ad}}(x)}{\sum_{y \in Q_t} f_{\text{ad}}(y)}$$
Luego se seleccionan aleatoriamente las parejas de padres. A esta forma de seleccionar se le conoce como método de la ruleta. Es importante mencionar que la función de adaptación se toma como la función objetivo a optimizar.

4. **Recombinación o cruzamiento:** La recombinación se efectúa para cada pareja de padres con probabilidad $p_c \in [0, 1]$ con el fin de generar dos descendientes donde se combinan las características de ambos padres.

5. **Mutación:** Se realiza para cada hijo y para cada gen con probabilidad $p_m \in [0, 1]$, se cambia aleatoria y uniformemente el gen por otro valor distinto del alfabeto. Esto per-

mite alcanzar zonas del espacio de búsqueda que no estaban cubiertas por los individuos de la población actual, en otras palabras, evita caer en óptimos locales.

6. **Reemplazo:** En el caso de los algoritmos genéticos canonicos, se conserva la descendencia generada. Mientras que en los algoritmos genéticos elitistas, que son los que se utilizan en este proyecto ocurre lo siguiente:

   - Si $f_{\text{ad}}(Q_{t+1}) \geq f_{\text{ad}}(Q_t)$, conservamos la descendencia.
   - Si $f_{\text{ad}}(Q_{t+1}) < f_{\text{ad}}(Q_t)$, reemplazamos al peor individuo por el mejor individuo obtenido hasta el momento.

   Donde para Q población, se ha definido:

$$f_{\text{ad}}(Q) = \max_{x \in Q} f_{\text{ad}}(x)$$

7. **Condición de término:** Es un criterio o condición de término, en general se usan dos: correr el algoritmo un número máximo de iteraciones (generaciones) o detenerlo cuando no haya cambios en la población. Mientras no se cumpla la condición de término, el algoritmo reinicia el proceso a partir de la etapa "selección", generando consigo una secuencia de poblaciones $(Q_t)_t$. Una vez que se alcanza cumple el criterio de parada, el algoritmo retorna una población final Q_T. De aquí se extrae la solución del problema:

$$\bar{x} = \text{argmáx}\{f_{\text{ad}}(x) : x \in Q_T\}$$

2.9. Estado del Arte en el Cálculo de Baricentros

La noción del baricentro de Wasserstein fue introducida por primera vez por [1] y luego ha sido investigada extensamente tanto teórica como numéricamente [[26], [27], [28], [13]]. Sin embargo, el cálculo de este objeto presenta desafíos, especialmente en entornos de alta dimensionalidad y continuos [4]. Los primeros estudios existentes sobre el cálculo del baricentro de Wasserstein de distribuciones continuas requieren la discretización de todas las distribuciones de entrada o del propio baricentro, lo cual no escala bien en entornos de alta dimensionalidad. Además, la técnica de discretización del baricentro continuo no es deseable, ya que carece de la naturaleza inherentemente continua de las distribuciones de datos y de la capacidad de generar nuevas muestras cuando sea necesario. La mayoría de estos estudios resuelven principalmente el problema del baricentro de Wasserstein discreto, es decir, promediando algunas distribuciones discretas, mediante programas lineales (o algún problema equivalente) [[29], [30], [31], [32]] o métodos basados en proyección regularizada [[33], [34], [35]] Sin embargo, estos métodos no pueden resolver el problema del baricentro de Wasserstein continuo, es decir, promediar algunas distribuciones continuas. Para abordar este problema, [[36], [37], [38]] asumen que el baricentro es un conjunto de puntos finitos y luego operan con las distribuciones de entrada continuas. Basados en algoritmos de transporte óptimo semi-discretos, encuentran una solución y utilizan la distribución discreta obtenida como aproximación del verdadero baricentro.

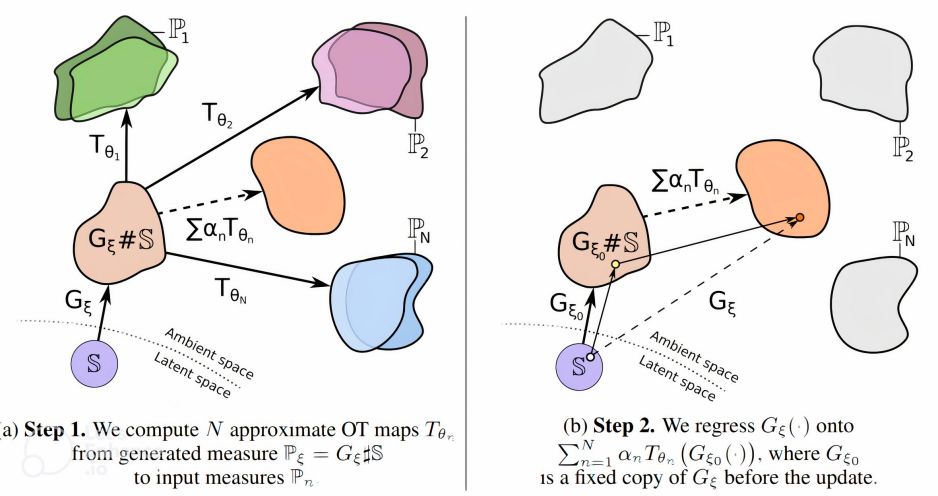
Los métodos de aproximación para los baricentros de Wasserstein de distribuciones de probabilidad continuas permanecían inexplorados hasta hace poco [[39], [40], [41]]. Todos ellos

se basan en la formulación dual del problema de baricentro de Wasserstein y representan los potenciales duales con redes neuronales. Sin embargo, estos métodos no son end-to-end (consisten en dos pasos secuenciales) y recuperan los baricentros a través de una estimación adicional. El método en [39] calcula los potenciales duales basados en una formulación dual regularizada mediante un algoritmo estocástico y recupera el baricentro utilizando los gradientes de los potenciales o mediante proyección baricentrica. Además, como señala [39], este método asume un prior fijo como estimación del baricentro desde el principio, lo que puede ser una mala aproximación para el verdadero baricentro y resultar en aproximaciones inexactas. Basados en el costo de base específico, es decir, $c(x, y) = \|x - y\|_2^2$, [40], [41] calculan los baricentros bajo la distancia de Wasserstein-2. El método en [40] utiliza una red generativa para representar un baricentro y lo recupera mediante un modelo generativo, lo cual sufre de las limitaciones habituales de las redes generativas como el *modus collapsis*. Aunque este método no involucra los términos de regularización, termina con un desafiante problema de min-max-min.

El método en [41] construye una formulación regularizada para asegurar que las funciones potenciales óptimas sean consistentes con el verdadero baricentro, y luego recupera el baricentro utilizando los gradientes de los potenciales como operador push-forward. Sin embargo, la regularización propuesta requiere la selección de una distribución previa que esté acotada por debajo del verdadero baricentro, lo cual es una tarea no trivial. Además, [42] propone un algoritmo genérico para calcular baricentros con respecto a discrepancias arbitrarias, que también parametriza el baricentro utilizando una red generativa. En [43] se mencioan que el desafío clave del problema del baricentro continuo de Wasserstein consiste en encontrar una representación adecuada del baricentro. Motivados por esto, se introduce una familia de distribuciones continuas, denominadas distribuciones variacionales, y se encuentra la más cercana como aproximación del verdadero baricentro. Así, se enmarca el problema de cálculo del baricentro como un problema de optimización, donde los parámetros de las distribuciones variacionales se ajustan para que sea similar al verdadero baricentro. Siguiendo este enfoque, se propone un método basado en una nueva formulación dual regularizada, donde se emplea regularización de monotonía cíclica c para obtener aproximaciones precisas. Desde el punto de vista teórico, se demuestra la convergencia del método propuesto.

2.9.1. WIN

El algoritmo desarrollado en [44] está basado en el algoritmo de punto fijo para el baricentro de Wasserstein. Lo que hacen es emplear un modelo generativo para parametrizar el baricentro, es decir, $\mu_\xi = G_\xi \# S$, donde S es una medida latente, por ejemplo, $S = N(0, I_H)$, y G_ξ es una red neuronal $\mathbb{R}^h \to \mathbb{R}^d$ donde $h << d$ con parámetros ξ. Para calcular el operador $H(P_\xi)$ y actualizar G_ξ consta de dos pasos, tal como se muestra en la figura 2.6. Primero, se aproximan las N transformaciones T_n tales que $T_n \# \mu_\xi = \nu_n$ mediante el uso de N redes $\{T_{\theta_n}, v_{\omega_n}\}$ y se entrenan optimizando (2.18) con $P \leftarrow \mu_\xi$ y $Q \leftarrow \nu_n$. Para cada $n = 1, 2, ..., N$, se usa SGD utilizando lotes de $G_\xi \# S$ y ν_n y con ello obtener $T_{\theta_n} \approx T_{\mu_\xi \to \nu_n}$. Segundo, la actualización de G_ξ para representar $H(\mu_\xi)$ en lugar de μ_ξ se hará a través de una regresión. Sea G_{ξ_0}, una copia fija de G_ξ. A continuación, se aproxima $G_\xi(\cdot)$ a $\sum_{n=1}^N \lambda_n T_{\theta_n}(G_{\xi_0}(\cdot))$ mediante regresión usando el error cuadrático medio. Los detalles de cómo se entrena cada red quedan resumidos en el siguiente pseudocódigo:



(a) **Step 1.** We compute N approximate OT maps T_{θ_n} from generated measure $\mathbb{P}_\xi = G_\xi \sharp \mathbb{S}$ to input measures $\mathbb{P}_n$.

(b) **Step 2.** We regress $G_\xi(\cdot)$ onto $\sum_{n=1}^N \alpha_n T_{\theta_n}(G_{\xi_0}(\cdot))$, where G_{ξ_0} is a fixed copy of G_ξ before the update.

Figura 2.6: Izquierda: primer paso del algoritmo. Derecha: segundo paso del algoritmo (imagen obtenida de [44])

Algorithm 2 Redes Iterativas de Wasserstein (WIN) para la Estimación del Baricentro

Require: Medida latente S y medidas de entrada $\nu_1, \ldots, \nu_N$; pesos $\lambda_1, \ldots, \lambda_N > 0$ ($\sum_{n=1}^N \lambda_n = 1$); número de iteraciones por red: K_G, K_T, K_v; generador $G_\xi : \mathbb{R}^h \to \mathbb{R}^d$; redes de mapeo $T_{\theta_1}, \ldots, T_{\theta_N} : \mathbb{R}^d \to \mathbb{R}^d$; potenciales $v_{\omega_1}, \ldots, v_{\omega_N} : \mathbb{R}^d \to \mathbb{R}$; regresión $l : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^+$;

Ensure: Generador G_ξ que satisface $G_\xi \# S \approx \mu$; mapas OT que satisfacen $T_{\theta_n} \# (G_\xi \# S) \approx \nu_n$.

repeat
  OT solver
  for $n = 1, 2, \ldots, N$ **do**
    for $K_v = 1, 2, \ldots, K_v$ **do**
      Muestrear lotes $Z \sim S$, $Y \sim \nu_n$. $X \leftarrow G_\xi(Z)$.
      $\mathcal{L}_v \leftarrow \frac{1}{|X|} \sum_{x \in X} v_{\omega_n}(T_{\theta_n}(x)) - \frac{1}{|Y|} \sum_{y \in Y} v_{\omega_n}(y)$.
      Actualizar ω_n utilizando el gradiente de $\mathcal{L}_v$.
    end for
    for $K_T = 1, 2, \ldots, K_T$ **do**
      Muestrear lotes $Z \sim S$ $X \leftarrow G_\xi(Z)$.
      $\mathcal{L}_T \leftarrow \frac{1}{|X|} \sum_{x \in X} \frac{1}{2} \|x - (T_{\theta_n}(x)\|^2 - v_{\omega_n}(T_{\theta_n}(x))$.
      Actualizar θ_n utilizando el gradiente de $\mathcal{L}_T$.
    end for
  end for
  Generator update
  $G_{\xi_0} \leftarrow G_\xi$
  for $K_G = 1, 2, \ldots, K_G$ **do**
    Muestrear lotes $Z \sim S$
    $\mathcal{L}_T \leftarrow \frac{1}{|Z|} \sum_{z \in Z} l(G_\xi(z), \sum_{n=1}^N \lambda_n T_{\theta_n}(G_{\xi_0}(z)))$.
  end for
until condición de parada

La figura 2.7 ilustra los resultados obtenidos por este algoritmo

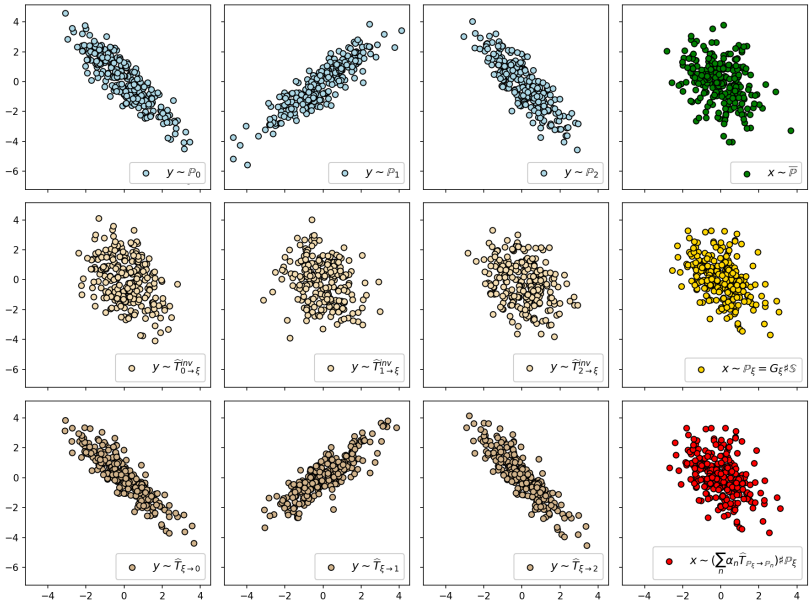


Figura 2.7: Baricentro de Wasserstein entre 3 gaussianas 2 dimensional.

La primera fila corresponde a las medidas input, y el gráfico de la esquina superior derecha indica el baricentro teórico. La segunda fila corresponde a los mapeos desde cada medida input hacia el baricentro. El último gráfico de esa fila indica la estimación del baricentro realizada por el algoritmo. La última fila muestra los mapeos desde el baricentro hacia cada medida input. Y el último gráfico es el operador de punto fijo.

Capítulo 3

Resultados

Los resultados teóricos de esta investigación se presentan a continuación. En primer lugar se formula el baricentro débil reverso de Wasserstein, sus propiedades matemáticas, ventajas y desventajas con respecto a los otros dos baricentros. En segundo lugar se diseñan e implementan algoritmos para resolver numéricamente el baricentro propuesto y el débil dado en [3].

3.1. Formulación del Problema

Siguiendo las formulaciones de los baricentros mencionados en el Marco Teórico, se propone definir de manera análoga el baricentro débil reverso

Definición 3.1 *Sean* $\{\nu_i\}_{i=1,\dots,n} \in \mathcal{P}_2(\mathbb{R}^d)$ *, pesos* $\lambda_1, \dots, \lambda_n$ *tales que* $\sum_{i=1}^n \lambda_i = 1$ *y* $M \in \mathbb{R}_+$

$$\min_{\mu \in \mathcal{F}_M} \sum_{i=1}^{n} \lambda_i V(\nu_i|\mu) = \min_{\mu \in \mathcal{F}_M} F(\mu) \tag{3.1}$$

Donde $\mathcal{F}_M = \{\mu \in \mathcal{P}_2(\mathbb{R}^d) \; ; \; \int_{\mathbb{R}^d} \|x\|^2 d\mu(x) \leq M\}$*. y* $F(\mu) = \sum_{i=1}^n \lambda_i V(\nu_i|\mu)$

De acuerdo al Teorema 2.2.7, lo que se esta haciendo es buscar una medida μ que promedia las medidas η_i más cercanas a cada ν_i según W_2^2 y que sea mayor en orden convexo que todas ellas. Aquí se aprecia la primera diferencia con respecto al baricentro débil. Es necesario imponer alguna restricción sobre la solución, pues ocurre que los momentos de orden 2 de μ no están acotados de forma natural por el problema pues ninguna medida involucrada en el cálculo de μ la mayora en orden convexo, lo que provoca que pueda no existir solución.

3.2. Proposiciones

A continuación se presentan diferentes propiedades que posee el baricentro débil reverso. La primera proposición garantiza que la formulación anterior tiene sentido.

Proposición 3.2.1. *El problema de baricentro reverso admite un minimizador* $\mu \in \mathcal{F}_M \; \forall M \in \mathbb{R}_+$.

Demostración. Sea $\mu_m \subseteq \mathcal{F}_M$ una sucesión minimizante. Luego

$$\begin{aligned}\sup_m \int \|x\|^2 d\mu_m &\leq \sup_m M \\ &= M \\ &< \infty\end{aligned}$$

Lo que prueba que la sucesión de medidas es tensa. Por el teorema de Prokhorov $\exists \mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ tal que una subsucesión μ_m converge débil a μ^*. Así, el conjunto $\mathcal{F}_M$ es relativamente compacto en la topología débil. Además, por Fatou para medidas se tiene que:

$$\begin{aligned}\int \|x\|^2 d\mu^*(x) &\leq \liminf \int \|x\|^2 d\mu_m(x) \\ &\leq \liminf M \\ &= M\end{aligned}$$

Por lo que $\mu^* \in \mathcal{F}_M$. Por lo tanto, este conjunto es compacto en la topología débil. Dado que

$$\sup_m \int \ \|x\|^2 d\mu_n(x) < \infty$$

por [45] se sabe que el conjunto $K = \{\mu_n\}_{n\in\mathbb{N}}$ tiene momento de orden ρ integrable $\forall \rho \in (0,2)$. En particular, para $\rho = 1$ se concluye que K tiene momentos 1 integrables. Gracias a la definición 1.10 se tiene que la sucesión μ_n converge débil en $\mathcal{P}_1$. Por el Teorema 1.2.11 μ_n converge en W_1, es decir, $W_1(\mu_n, \mu^*) \to 0$. En consecuencia, el conjunto $\mathcal{F}_M$ es compacto para la topología dada por W_1. Por el Teorema 2.2.8 se tiene la continuidad de $V(\nu_i|\cdot)$ en W_1, por ende la función $F(\mu)$ es W_1 continua. Como se tiene una función continua en un compacto, existe mínimo. □

Proposición 3.2.2. *Si μ es un baricentro débil reverso de $\{\nu_i\}_{i=1}^n \in \mathcal{P}_2(\mathbb{R}^d)$ y $\mu' \in \mathcal{F}_M$ es tal que $\mu \leq_c \mu'$ Entonces μ' también es un baricentro débil reverso.*

Demostración. Sea $A = \{\eta \in \mathcal{P}_2(\mathbb{R}^d) | \eta \leq_c \mu\}$, $B = \{\eta \in \mathcal{P}_2(\mathbb{R}^d) | \eta \leq_c \mu'\}$ por hipótesis más la transitividad del orden convexo se tiene que $A \subseteq B$ y usando la proposición: Si $A \subseteq B \Rightarrow \inf_A \geq \inf_B$ se tiene lo siguiente:

$$\begin{aligned}V(\nu_i|\mu) &= \inf_{\eta\in A} W_2^2(\nu_i, \eta) \\ &\geq \inf_{\eta\in B} W_2^2(\nu_i, \eta) \\ &= V(\nu_i|\mu')\end{aligned}$$

Lo anterior es $\forall i = 1, \ldots, n$ por ende $F(\mu) \geq F(\mu')$ pero como μ es solución del problema, se concluye que $F(\mu) = F(\mu')$ □

Las siguientes proposiciones buscan encontrar una solución explicita en función de las medidas $\{\nu_i\}_{i=1}^n$ en el caso cuando están centradas. Para ello se estudia qué relación tiene la convolución de las ν_i con cada una de ellas.

Proposición 3.2.3. *Sean $Y_i \sim \nu_i \in \mathcal{P}_2(\mathbb{R}^d), i \in \{1, \dots, n\}$ v.as independientes, y $\hat{\nu}_i$ su versión centrada. Se tiene que $*_{i=1}^n \hat{\nu}_i \geq_c \hat{\nu}_i \quad \forall i$*

Demostración. Sea ϕ una función convexa cualquiera, por inducción

- Caso base. Sean X,Y 2 v.a. centradas e independientes. Luego se tiene que

$$\begin{aligned}\mathbb{E}(\phi(X+Y)) &= \mathbb{E}(\mathbb{E}(\phi(X+Y)|X)) \\ &= \mathbb{E}(\mathbb{E}(\phi(Y+x)|_{X=x})) \\ &\geq \mathbb{E}(\phi(\mathbb{E}(x+Y)|_{X=x})) \text{ Jensen} \\ &= \mathbb{E}(\phi(X)) \quad \mathbb{E}(Y) = 0\end{aligned}$$

  Y como la convolución es la ley de la suma de variables aleatorias independientes, se concluye.

- Paso inductivo. Basta aplicar el caso base con $X = X_{n+1}$ e $Y = \sum_{i=1}^n X_i$ y descomponer iterativamente hasta llegar al i que quiero. Este procedimiento se puede repetir para todo i. Por lo tanto $\mathbb{E}(\phi(\sum_{i=1}^n X_i)) \geq \mathbb{E}(\phi(X_i))$

□

Proposición 3.2.4. *Sea $\mu \in \mathcal{F}_M$. Entonces $\mu \geq_c \nu_i$, $\forall i$ tal que $\lambda_i > 0$ ssi $F(\mu) = 0$*

Demostración. $\Rightarrow$ Por hipótesis se sabe que $V(\nu_i|\mu) = 0 \; \forall i$
$\Leftarrow$ Si F(u)=0 entonces $\forall i \;\; \inf_{\eta \leq_c \mu} W_2^2(\nu_i, \eta) = 0$ equivale a decir que $\forall i \; \exists \eta_i \; W_2^2(\eta_i, \nu_i) = 0$ y $\eta_i \leq_c \mu$. Como W es distancia $\eta_i = \nu_i$ asi $\nu_i \leq_c \mu$ □

Proposición 3.2.5. *Sean $Y_i \sim \nu_i \in \mathcal{P}_2(\mathbb{R}^d), i \in \{1, \dots, n\}$ v.as independientes, y $\hat{\nu}_i$ su versión centrada. Sea $M \in \mathbb{R}_+$, se tiene que si $*_{i=1}^n \hat{\nu}_i \in \mathcal{F}_M$ entonces es solución del problema de baricentro reverso para $\hat{\nu}_i$.*

Demostración. Por la proposición 3.2.4 se sabe que que $*_{i=1}^n \hat{\nu}_i \geq_c \hat{\nu}_i \; \forall i$, por ende $V(\nu_i| *_{i=1}^n \hat{\nu}_i) = 0$. Como $F(u) \geq 0$ y $*_{i=1}^n \hat{\nu}_i \in \mathcal{F}_M$ entonces $*_{i=1}^n \hat{\nu}_i$ es solución del problema.

□

Tal como en el baricentro débil, en el baricentro reverso existe una solución que se puede computar directamente conociendo las medidas input involucradas. La principal diferencia es que en el caso reverso se puede escoger un M lo suficientemente pequeño tal que $*_{i=1}^n \hat{\nu}_i \notin \mathcal{F}_M$, y por la proposición 3.2.3 el problema tiene solución $\forall M \in \mathbb{R}_+$ lo cual provoca que se puedan obtener soluciones más interesantes que la convolución en cualquier contexto. Cabe destacar que en el caso débil no se puede realizar esto, por ende siempre se podría obtener solo la masa de dirac como solución, siendo quizás poco relevante en la práctica. Otra diferencia es que el cálculo de la solución conocida para el baricentro reverso solo se puede realizar en el caso de que las medidas input estén centradas y sean leyes de variables aleatorias independientes, siendo que en su contraparte se puede computar la masa de dirac para cualquier medida considerada.

Cabe destacar que, tal como el baricentro débil en el caso de distribuciones Gaussianas unidimensionales $\nu_i = \mathcal{N}(m, \sigma_i^2)$, el conjunto de baricentros débiles incluye $\{\mu = \mathcal{N}(m, \sigma^2) \mid 0 \leq$

$\sigma^2 \leq \min_{1\leq i\leq n} \sigma_i^2\}$. En el caso reverso, el conjunto de baricentros débiles reversos incluye $\{\mu = \mathcal{N}(m, \sigma^2) \mid \max_{1\leq i\leq n} \sigma_i^2 \leq \sigma^2 \leq M\}$ siempre y cuando $\max_{1\leq i\leq n} \sigma_i^2 < M$. En caso contrario, hay que buscar la mayor varianza dentro de las n-1 restantes que sea menor estricta a M.

Ya que se tiene la solución explicita para M suficientemente grande en el caso ν_i centradas, la siguiente pregunta a responder es como se comporta el baricentro débil reverso cuando las ν_i no están centradas.

Proposición 3.2.6. *Sean $Y_i \sim \nu_i, i \in \{1, \ldots, n\}$ v.as y $X \sim \mu$. Si μ es baricentro débil reverso para ν_i entonces $\hat{\mu}$, la ley de la versión centrada de X, es baricentro débil reverso para $\hat{\nu}_i$*

Demostración.

$$\begin{aligned} V(\nu_i|\mu) &= \inf_{\eta\leq_c \mu} W_2^2(\nu_i, \eta) \\ &= \inf_{\eta\leq_c \hat{\mu}} W_2^2(\hat{\nu}_i, \eta) + \|\mathbb{E}_\mu(X) - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\ &= V(\hat{\nu}_i|\hat{\mu}) + \|\mathbb{E}_\mu(X) - \mathbb{E}_{\nu_i}(Y_i)\|^2 \end{aligned}$$

Lo anterior dice que minimizar $\sum_{i=1}^n V(\nu_i|\mu)$ sobre $\mathcal{F}_M$ es equivalente a minimizar $\sum_{i=1}^n \lambda_i V(\nu_i|\mu) + \sum_{i=1}^n \lambda_i \|\omega - \mathbb{E}_{\nu_i}(Y_i)\|^2$ sobre 2 parámetros independientes $\hat{\mu} \in \mathcal{F}_M$ centrada y $\omega \in \mathbb{R}^n$, tomando μ como la ley de $X = X' + \omega$, $X' \sim \hat{\mu}$ gracias a que la varianza de una variable aleatoria es invariante bajo traslaciones, esto quiere decir que $\forall \omega \in \mathbb{R}^n$ X, X' tienen el mismo momento de orden 2.

Si $\hat{\mu}$ no fuese baricentro débil reverso para $\hat{\nu}_i$ se tendría que existe $\mu_0 \in \mathcal{F}_M$ tal que

$$\min_{\mu\in\mathcal{F}_M} \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\mu) = \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\mu_0) < \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu})$$

y con ello

$$\begin{aligned} \min_{\mu\in\mathcal{F}_M} \sum_{i=1}^n \lambda_i V(\nu_i|\mu) &= \sum_{i=1}^n \lambda_i V(\nu_i|\mu) \\ &= \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}) + \|\mathbb{E}_\mu(X) - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\ &> \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\mu_0) + \|\mathbb{E}_\mu(X) - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\ &\geq \min_{(\hat{\mu}_1,\omega)\in\mathcal{F}_M\times\mathbb{R}^n} \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}_1) + \|\omega - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\ &= \min_{\mu\in\mathcal{F}_M} \sum_{i=1}^n \lambda_i V(\nu_i|\mu) \\ &= \sum_{i=1}^n \lambda_i V(\nu_i|\mu) \end{aligned}$$

con lo que se concluye que $\sum_{i=1}^n \lambda_i V(\nu_i|\mu) > \sum_{i=1}^n \lambda_i V(\nu_i|\mu)$ lo cual es una contradicción. Por lo tanto $\hat{\mu}$ es baricentro débil reverso para $\hat{\nu}_i$ □

Proposición 3.2.7. *Sean $Y_i \sim \nu_i, i \in \{1,\dots,n\}$ v.as. Si $X' \sim \hat{\mu}$ es baricentro débil reverso para $\hat{\nu}_i$ entonces la ley μ de la variable aleatoria $X = X' + \omega$ es baricentro débil reverso para ν_i, donde $\omega = \arg\min_{\omega\in\mathbb{R}^n} \sum_{i=1}^n \lambda_i \|w - \mathbb{E}_{\nu_i}(Y_i)\|^2$*

Demostración. Sea μ_0 baricentro debil reverso para ν_i se tiene que

$$\begin{aligned}
\sum_{i=1}^n \lambda_i V(\nu_i|\mu_0) &= \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}_0) + \|\mathbb{E}_\mu(X) - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&\geq \min_{(\hat{\mu}_1,\omega_1)\in\mathcal{F}_M\times\mathbb{R}^n} \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}_1) + \|\omega_1 - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&= \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}) + \|\omega - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&= \sum_{i=1}^n \lambda_i V(\nu_i|\mu)
\end{aligned}$$

y como μ_0 es baricentro débil reverso, entonces $\sum_{i=1}^n \lambda_i V(\nu_i|\mu_0) \leq \sum_{i=1}^n \lambda_i V(\nu_i|\mu)$ por ende μ tiene el mismo valor objetivo que μ_0 y por lo tanto es un baricentro débil reverso para ν_i. □

Observación 3.2.8. Las dos proposiciones anteriores dicen que para obtener baricentros en cualquier caso, independiente de la elección de M, basta con computar en un caso y después trasladarlo según corresponda. Por ejemplo, para encontrar una solución en el caso ν_i no centrado, basta con centrarlas y computar la solución explícita, la cual está dada por la convolución de las ν_i, y posteriormente trasladar el baricentro obtenido según el ω explicitado por la proposición 3.2.7. Es importante destacar que el ejemplo anterior es válido siempre y cuando se tenga un M lo suficientemente grande para que la convolución pertenezca a $\mathcal{F}_M$.

Observación 3.2.9. La media del baricentro reverso cumple que $\mathbb{E}_\mu(X) = \sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i}(Y_i)$. Además, se puede decir que para M suficientemente grande μ es baricentro débil reverso para ν_i independientes sí y solo sí $\hat{\nu}_i \leq \hat{\mu} \quad \forall i$. Más aún, gracias a esto y al hecho que $V(\nu_i|\mu) = 0$ ssi $\nu_i \leq_c \mu$ se puede concluir que

$$\begin{aligned}
\min_{\mu\in\mathcal{F}_M} \sum_{i=1}^n \lambda_i V(\nu_i|\mu) &= \min_{(\hat{\mu},\omega)\in\mathcal{F}_M\times\mathbb{R}^n} \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}) + \|\omega - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&= \min_{\hat{\mu}\in\mathcal{F}_M} \sum_{i=1}^n \lambda_i V(\hat{\nu}_i|\hat{\mu}) + \min_{\omega\in\mathbb{R}^n} \sum_{i=1}^n \|\omega - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&= \min_{\omega\in\mathbb{R}^n} \sum_{i=1}^n \lambda_i \|\omega - \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&= \sum_{i=1}^n \lambda_i \|\mathbb{E}_{\nu_i}(Y_i)\|^2 - \|\sum_{i=1}^n \lambda_i \mathbb{E}_{\nu_i}(Y_i)\|^2 \\
&= \min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i V(\mu|\nu_i)
\end{aligned}$$

Lo anterior señala que el valor del problema de optimización asociado a ambos baricentros tiene el mismo valor objetivo.

La observación 3.2.9 otorga el primer indicio de que los baricentros débiles poseen relaciones entre ellos. Una pregunta esencial es que tan lejos estan entre sí, y una primera aproximación a esta respuesta esta dada por la siguiente proposición.

Proposición 3.2.10. *Sea μ_0 baricentro débil y μ_1 baricentro débil reverso. Entonces*

$$W_2^2(\mu_0, \mu_1) \leq 2\sum_{i=1}^{n} \lambda_i \int \|y\|^2 d\nu_i(y) + 2M$$

Demostración. Por el colorario 2.4.7 se sabe que el baricentro débil verifica que $\mu_0 = \left(\sum_{i=1}^n \lambda_i S_{\mu_0}^{\nu_i}(x)\right) \# \mu_0$

$$\begin{aligned} W_2^2(\mu_0, \mu_1) &\leq \int\int \|x-y\|^2 d\mu_0(x) d\mu_1(y) \\ &= \int\int \|\sum_{i=1}^{n} \lambda_i S_{\mu_0}^{\nu_i}(x) - y\|^2 d\mu_0(x) d\mu_1(y) \\ &\leq 2\int \|\sum_{i=1}^{n} \lambda_i S_{\mu_0}^{\nu_i}(x)\|^2 d\mu_0(x) + 2M \\ &\leq 2\sum_{i=1}^{n} \lambda_i \int \|y\|^2 d\nu_i(y) + 2M \end{aligned}$$

□

Proposición 3.2.11. *Sean $\{\nu_i\}_{i=1}^n \subseteq \mathcal{P}_2(\mathbb{R}^d)$ centradas. μ_0, μ_1 baricentro débil y reverso respectivamente. si existe i_0 tal que $\nu_{i_0} \leq_c \mu_1$ entonces $\mu_0 \leq \mu_1$*

Demostración. Gracias al Teorema 2.4.5 se tiene que $\mu_0 \leq_c \nu_i \ \forall i$ y por hipótesis existe i_0 tal que $\nu_{i_0} \leq_c \mu_1$ por la transitividad del orden convexo entre medidas se concluye el resultado. □

Observación 3.2.12. La hipótesis de que existe i_0 tal que $\nu_{i_0} \leq_c \mu_1$ se puede omitir para M suficientemente grande

Los siguientes Teoremas otorgan la principal relación entre el baricentro débil reverso y las medidas input

Teorema 3.2.13. *Sean $Y_i \sim \nu_i, i \in \{1, \ldots, n\}$ v.as independientes y $X \sim \mu$ baricentro débil reverso de ν_i. Sea M suficientemente grande (por ejemplo M tal que la convolución de medidas pertenece a $\mathcal{F}_M$) entonces*

$$X = Y_i - (\mathbb{E}(Y_i) - \mathbb{E}(X)) + \tilde{X}_i$$

donde $\tilde{X}_i = X - \mathbb{E}(X|Y_i)$. Finalmente, se tiene que $\mathbb{E}(X - \mathbb{E}(X)|Y_i - \mathbb{E}(Y_i)) = Y_i - \mathbb{E}(Y_i)$ o de manera equivalente $\hat{\nu}_i \leq_c \hat{\mu}$. Con $\hat{\nu}_i$, $\hat{\mu}$ las leyes de $Y_i - \mathbb{E}(Y_i)$ y $X - \mathbb{E}(X)$ respectivamente.

Demostración. Por Strassen + Teorema 2.2.7

- $(Y, X) \sim \pi^{\nu,\mu}$, $X \sim \mu$, $Y \sim \nu$
- $Z = S_\nu^\mu(Y) = \mathbb{E}(X|Y)$ *c.s* tiene ley η^* y es *coupling* óptimo con Y.

- (Z,X) es martingala, $\mathbb{E}(X|Z) = Z$ $c.s$

Además, sea $\tilde{X} \sim \tilde{\mu}$ baricentro débil. En este caso se tiene algo similar.

Por Strassen + Teorema 2.2.7

- $(\tilde{X}, Y) \sim \pi^{\tilde{\mu},\nu}$, $\tilde{X} \sim \tilde{\mu}$, $Y \sim \nu$
- $\tilde{Z} = S^{\nu}_{\tilde{\mu}}(\tilde{X}) = \mathbb{E}(Y|\tilde{X})$ $c.s$ tiene ley η_0^* y es *coupling* óptimo con Y.
- $(\tilde{Z}, Y)$ es martingala, $\mathbb{E}(Y|\tilde{Z}) = \tilde{Z}$ $c.s$

$X = S^{\mu}_{\nu}(Y) + X - \mathbb{E}(X|Y)$ Por la observación 3.2.9 se sabe que

$$\sum_{i=1}^{n} \lambda_i V(\nu_i|\mu) = \sum_{i=1}^{n} \lambda_i V(\tilde{\mu}|\nu_i) = \sum_{i=1}^{n} \lambda_i V(\delta_{\mathbb{E}(\tilde{X})}|\nu_i) \tag{3.2}$$

Luego

$$\begin{aligned}
V(\nu_i|\mu) =& W_2^2(\nu_i, \eta^*) \\
=& \mathbb{E}\|Y_i - Z_i\|^2 \\
\geq& \|\mathbb{E}Y_i - \mathbb{E}Z_i\|^2 \\
=& \|\mathbb{E}\tilde{Z}_i - \mathbb{E}X\|^2 \text{ Pues } \mathbb{E}Y_i = \mathbb{E}\tilde{Z}_i \text{ y } \mathbb{E}X = \mathbb{E}Z_i \\
=& \|\mathbb{E}\tilde{X} - \mathbb{E}\tilde{Z}_i\|^2 \text{ Pues } \mathbb{E}X = \mathbb{E}\tilde{X} \\
=& W_2^2(\delta_{\mathbb{E}\tilde{X}}, \delta_{\tilde{Z}_i}) \\
\geq& \inf_{\eta \leq_c \nu_i} W_2^2(\delta_{\mathbb{E}\tilde{X}}, \eta) = V(\delta_{\mathbb{E}(\tilde{X})}|\nu_i)
\end{aligned}$$

Por lo anterior y junto a (3.2) se tiene que $\forall i$ $V(\nu_i|\mu) = V(\tilde{\mu}|\nu_i) = V(\delta_{\mathbb{E}(\tilde{X})}|\nu_i)$. Además, todas las desigualdades anteriores se convierten en igualdad, por lo tanto $\mathbb{E}\|Y_i - Z_i\|^2 = \|\mathbb{E}Y_i - \mathbb{E}Z_i\|^2$ esto implica que $Y_i - Z_i$ es determinista. En consecuencia $Y_i - Z_i = \mathbb{E}Y_i - \mathbb{E}Z_i$ y como $\mathbb{E}Z_i = \mathbb{E}X$ se tiene que $Z_i = Y_i - (\mathbb{E}Y_i - \mathbb{E}X)$.

$$\begin{aligned}
X =& Y_i - (\mathbb{E}Y_i - \mathbb{E}X) + X - \mathbb{E}(X|Y_i) \\
=& Y_i - (\mathbb{E}Y_i - \mathbb{E}X) + \tilde{X}_i
\end{aligned}$$

Finalmente, como $Z_i = Y_i - (\mathbb{E}Y_i - \mathbb{E}X)$ se tiene que

$$\begin{aligned}
\mathbb{E}(X|Y_i) - \mathbb{E}X &= Y_i - \mathbb{E}Y_i \\
\mathbb{E}(X - \mathbb{E}(X)|Y_i) &= Y_i - \mathbb{E}Y_i
\end{aligned}$$

Como la función $p(x) = x - a$ es una función continua biyectiva con inversa continua, se tiene que la sigma algebra generada por Y_i y $Y_i - \mathbb{E}(Y_i)$ son las mismas pues los abiertos se preservan via $p(x)$. Por lo tanto $\mathbb{E}(X - \mathbb{E}(X)|Y_i) = \mathbb{E}(X - \mathbb{E}(X)|Y_i - \mathbb{E}(Y_i))$ y se concluye que $\mathbb{E}(X - \mathbb{E}(X)|Y_i - \mathbb{E}(Y_i)) = Y_i - \mathbb{E}Y_i$.

□

Observación 3.2.14. Notar que de la demostración del Teorema 3.2.13 se desprende que $\forall i\ V(\nu_i|\mu) = V(\tilde{\mu}|\nu_i) = V(\delta_{\mathbb{E}(\tilde{X})}|\nu_i)$

Este teorema garantiza que, para M suficientemente grande, y tanto en el caso centrado como no, se obtiene que todas las medidas input son variables latentes para el baricentro. El siguiente teorema da indicios de lo que sucede en la situación donde M no es lo suficientemente grande.

Teorema 3.2.15. *Sean $Y_i \sim \nu_i, i \in \{1, \dots, n\}$ v.as y $X \sim \mu$ baricentro débil reverso de ν_i. Se tiene que para cada i tal que $V(\nu_i|\mu) = 0$*

$$X = Y_i + \tilde{X}_i$$

donde $\tilde{X}_i = X - \mathbb{E}(X|Y_i)$.

Demostración. Por Strassen + Teorema 2.2.7

- $(Y, X) \sim \pi^{\nu,\mu}$, $X \sim \mu$, $Y \sim \nu$
- $Z = S^{\mu}_{\nu}(Y) = \mathbb{E}(X|Y)$ *c.s* tiene ley η^* y es *coupling* óptimo con Y.
- (Z,X) es martingala, $\mathbb{E}(X|Z) = Z$ *c.s*

$X = S^{\mu}_{\nu}(Y) + X - \mathbb{E}(X|Y)$ Además, se sabe que que $V(\nu_i|\mu) = 0$

$$\begin{aligned} 0 &= \inf_{\eta \leq_c \mu} W_2^2(\nu_i, \eta) \\ &\geq \inf_{\eta \leq_c \mu} \mathbb{E}\|Y_i - Z_i\|^2 \ Y_i,\ Z_i \text{ optimo } W_2 \text{ de } \nu_i \text{ y } \eta \\ &\geq \inf_{\eta \leq_c \mu} \|\mathbb{E}Y_i - \mathbb{E}Z_i\|^2 \geq 0 \end{aligned}$$

Por lo anterior se tiene que $\mathbb{E}\|Y_i - Z_i\|^2 = \|\mathbb{E}Y_i - \mathbb{E}Z_i\|^2$ esto implica que $Y_i - Z_i$ es determinista. Como $\mathbb{E}Z_i = \mathbb{E}X$, usando la igualdad obtenida se tiene que $Z_i = Y_i - (\mathbb{E}Y_i - \mathbb{E}X)$. Más aún, gracias a que $V(\nu_i|\mu) = 0$ se sabe que $\nu_i \leq_c \mu$ y al estar en relación de orden convexo esto implica necesariamente que $\mathbb{E}(X) = \mathbb{E}(Y_i)$. Por lo tanto:

$$\begin{aligned} X &= Y_i + X - \mathbb{E}(X|Y_i) \\ &= Y_i + \tilde{X}_i \end{aligned}$$

□

Teorema 3.2.16. *Sean $Y_i \sim \nu_i, i \in \{1, \dots, n\}$ v.as y $X \sim \mu$ baricentro débil reverso de ν_i. Se tiene que para cada i tal que $V(\mu|\nu_i) = 0$*

$$Y_i = X + \tilde{Y}_i$$

donde $\tilde{Y}_i = Y_i - \mathbb{E}(Y_i|X)$.

Demostración. Por Strassen + Teorema 2.2.7

- $(X, Y) \sim \pi^{\mu,\nu}$, $X \sim \mu$, $Y \sim \nu$

- $Z = S_\mu^\nu(X) = \mathbb{E}(Y|X)$ *c.s* tiene ley η^* y es *coupling* óptimo con X.
- (Z,Y) es martingala, $\mathbb{E}(Y|Z) = Z$ *c.s*

$Y_i = S_\mu^\nu(X) + Y_i - \mathbb{E}(Y_i|X)$ Además, se sabe que que $V(\mu|\nu_i) = 0$

$$\begin{aligned} 0 &= \inf_{\eta \leq_c \mu} W_2^2(\mu, \eta) \\ &\geq \inf_{\eta \leq_c \mu} \mathbb{E}\|X - Z_i\|^2 \ X,\ Z_i \text{ optimo } W_2 \text{ de } \mu \text{ y } \eta \\ &\geq \inf_{\eta \leq_c \mu} \|\mathbb{E}X - \mathbb{E}Z_i\|^2 \geq 0 \end{aligned}$$

Por lo anterior se tiene que $\mathbb{E}\|X - Z_i\|^2 = \|\mathbb{E}X - \mathbb{E}Z_i\|^2$ esto implica que $X - Z_i$ es determinista. Como $\mathbb{E}Z_i = \mathbb{E}Y_i$, usando la igualdad obtenida se tiene que $Z_i = X + \mathbb{E}Y_i - \mathbb{E}X$. Más aún, gracias a que $V(\mu|\nu_i) = 0$ se sabe que $\mu \leq_c \nu_i$ y al estar en relación de orden convexo esto implica necesariamente que $\mathbb{E}(X) = \mathbb{E}(Y_i)$. Por lo tanto:

$$\begin{aligned} Y_i &= X + Y_i - \mathbb{E}(Y_i|X) \\ &= X + \tilde{Y}_i \end{aligned}$$

□

El teorema anterior se puede interpretar como sigue: Dado $M \in \mathbb{R}_+$ cada $Y_i \sim \nu_i$ tal que $V(\nu_i|\mu) = 0$ es una variable latente al baricentro reverso, lo que equivale a decir que esta distribución contiene toda la información de la v.a Y_i. Más aún, si $V(\mu|\nu_i) = 0$ se puede decir que el baricentro reverso es una variable latente para Y_i, lo que implica que Y_i contiene toda la información del baricentro reverso.

Estos teoremas otorgan la principal diferencia entre ambos baricentros, el débil siempre corresponde a la distribución de una variable latente común a todas las distribuciones de entrada. En cambio en el caso reverso algunas medidas input podrían no tener relación alguna de este tipo con el baricentro, y no es claro cuanta información de aquellas v.as Y_i conserva. Todo depende de la elección de M.

Observación 3.2.17. Una pregunta interesante es ver cómo influye la elección del M en las soluciones. Esta constante tiene directa relación con el momento de orden 2 de las soluciones, y $\|x\|^2$ es una de las funciones convexas donde puede influir la elección de M. En otras palabras, M determina cuando el baricentro débil reverso y las ν_i no están en orden convexo. Por ejemplo, si se elige $M < \max_{1 \leq i \leq n} \sigma_i^2$ entonces para i_0 tal que $\sigma_{i_0}^2 = \max_{1 \leq i \leq n} \sigma_i^2$ no se puede tener que $\nu_i \leq_c \mu$ o si no $\sigma_{i_0} = \int \|x\|^2 d\nu_{i_0} \leq \int \|x\|^2 d\mu < M$ lo que contradice la elección de M. Lo anterior implica que $V(\nu_i|\mu) > 0$ lo que provoca que la función objetivo aumente su valor.

Sea $G(M) : \mathbb{R}_+ \to \mathbb{R}_+$ definida como $G(M) = \min_{\mu \in \mathcal{F}_\mathcal{M}} F(\mu)$ que queda bien definida gracias a la proposición 3.2.1. Otro factor es que si $M_1 < M_2$, entonces se tiene que $\mathcal{F}_{M_1} \subseteq \mathcal{F}_{M_2}$ y como $A \subseteq B \Rightarrow \inf_A \geq \inf_B$, entonces $G(M_2) \leq G(M_1)$. Por ende G es una función monótona decreciente, y por el Teorema de Lebesgue es derivable ctp. Además, el conjunto de puntos de discontinuidades de G puede contener a las varianzas de las ν_i, pues al tomar el

límite por la izquierda $\lim_{M\to\sigma_i^2-}$, en este caso es seguro que el baricentro reverso no estará por sobre ν_i en orden convexo, por tanto $V(\nu_i|\mu) > 0$. En cambio, cuando se considera el límite por la derecha $\lim_{M\to\sigma_i^2+}$, se tienen 2 casos:

1. $V(\nu_i|\mu) = 0$ aquí se presenta un punto de discontinuidad, pues el valor del problema va a incrementar al considerar M<σ_i^2.
2. $V(\nu_i|\mu) > 0$ no hay punto de discontinuidad.

3.3. Algoritmo para Baricentro Débil

Se propone fusionar los algoritmos WIN y NOT como enfoque. La premisa subyacente es aprovechar el hecho de que para el cálculo del baricentro débil existe un algoritmo de punto fijo, tal como se mostró en la Subsección 2.4.1. Además, en el caso débil, se pueden obtener las mismas garantías teóricas de reducción de la función a optimizar en cada iteración que en el caso no débil, según se describe en la Proposición 2.4.8. El algoritmo consiste en lo siguiente: se realiza una optimización en 2 pasos. La primera parte consiste en computar los *transport plan* entre el baricentro, el cual está parametrizado por una red neuronal G(Z) donde Z es una distribución no atómica, y las n medidas a promediar. Este cálculo se realiza siguiendo la misma metodología explicada para NOT en la sección 2.3. En otras palabras, se usará NOT para poder computar las funciones $T(x, z)$ que emulan los *transport plans* en el caso débil.

Para la segunda parte del algoritmo, es importante notar que la proyección baricéntrica se puede obtener de la siguiente manera:

$$S_\mu^\nu(x) = \int_Z T(x, z)dS(z) \tag{3.3}$$

. Se usará esto para modificar la regresión que se hace en WIN para encontrar los parámetros óptimos de la red G. Concretamente, se modifican las funciones $T(x)$ por los mapas estocásticos $T(x, z)$ y se estima la proyección baricéntrica mediante montecarlo con samples de la variable z. En otras palabras, se modifica el operador 2.1.1 por la función dada en 2.4.1. Todo lo anterior se resume en el siguiente pseudocódigo:

3.4. Algoritmo para Baricentro Débil Reverso Caso 1D

Con fines ilustrativos, se ocupará un algoritmo genético para poder computar el baricentro débil reverso en el caso gaussiano unidimensional, pues en este caso es donde se pueden manifestar de manera gráfica y numérica diferentes propiedades teóricas de este objeto que fueron discutidas en la sección de resultados.

La idea será plantear que el baricentro es una distribución gaussiana, y mediante un algoritmo genético se buscarán los parámetros óptimos de esta medida sujeto a que minimice la función objetivo (3.1). Para respetar la restricción asociada al problema de optimización, se creará una población inicial de varianzas que sean menor o igual a M. Gracias a como se realizan las diferentes operaciones de mutación, combinación, entre otras. Siempre se obtendrán

Algorithm 3 Redes Iterativas de Wasserstein (WIN) para la Estimación del Baricentro débil

Require: Medida latente S con soporte en $\mathbb{R}^h$ y $h << d$, U ruido gaussiano y medidas de entrada $\nu_1, \ldots, \nu_N$; pesos $\lambda_1, \ldots, \lambda_N > 0$ ($\sum_{n=1}^{N} \lambda_n = 1$); número de iteraciones por red: K_G, K_T, K_v; generador $G_\xi : \mathbb{R}^h \to \mathbb{R}^d$; redes de mapeo $T_{\theta_1}, \ldots, T_{\theta_N} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$; potenciales $f_{\omega_1}, \ldots, f_{\omega_N} : \mathbb{R}^d \to \mathbb{R}$; pérdida de regresión $l : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$;

Ensure: Generador G_ξ que satisface $G_\xi \# S \approx P$ baricentro; mapas estocásticos que representado los transport plans entre el baricentro y cada P_n.

repeat
  WOT solver
  for n =1,2,...,N **do**
    Muestrear lotes $Z \sim S$, $Y \sim \nu_n$. $X \leftarrow G_\xi(Z)$
    for cada $x \in X$ **do**
      Muestrear lote $Z_x \sim$U
      $\mathcal{L}_f \leftarrow \frac{1}{|X|} \sum_{x \in X} \frac{1}{|Z_x|} \sum_{z \in Z_x} f_\omega(T_{\theta_n}(x, z)) - \frac{1}{|Y|} \sum_{y \in Y} f_{\omega_n}(y)$
      Actualizar ω_n usando $\frac{\partial \mathcal{L}_f}{\partial \omega_n}$;
    end for
    for $k_T = 1, 2, \ldots, K_T$ **do**
      Muestrear batches $Z \sim S$, $X \leftarrow G_\xi(Z)$
      Para cada $x \in X$ muestrear batch $Z_x \sim \mathbb{U}$
      $\mathcal{L}_T \leftarrow \frac{1}{|X|} \sum_{x \in X} \hat{C}(x, T_{\theta_n}(x, z)) - \frac{1}{|Z_x|} \sum_{z \in Z_x} f_{\omega_n}(T_{\theta_n}(x, z))$
      Actualizar θ_n usando $\frac{\partial \mathcal{L}_T}{\partial \theta_n}$
  end for
end for
Generator update
$G_{\xi_0} \leftarrow G_\xi$
for $K_G = 1, 2, \ldots, K_G$ **do**
  Muestrear lotes $Z \sim S$
  for cada $z \in Z$ **do**
    Muestrear lote $Z_z \sim$U
    $\mathcal{L}_G \leftarrow \frac{1}{|Z|} \sum_{z \in Z} l(G_\xi(z), \sum_{n=1}^{N} \lambda_n \frac{1}{|Z_z|} \sum_{z_z \in Z_z} T_{\theta_n}(G_{\xi_0}(z), z_z))$.
  end for
end for
until condición de parada

individuos factibles.

Algorithm 4 Algoritmo Genético para estimar el baricentro reverso 1D

Require: medidas de entrada $\nu_1, \ldots, \nu_N$; pesos $\lambda_1, \ldots, \lambda_N > 0$ ($\sum_{n=1}^{N} \lambda_n = 1$), $M > 0$, $p_{cruz}, p_{mut}, p_{selec}$, N= número de generaciones ;

Ensure: parámetros ξ, σ^2 de una distribución Gaussiana 1D

1: Se genera una población inicial de medias y varianzas. Se crea una grilla de 11x11 con 11 medias y varianzas en el conjunto $[22.5, 67.5] \times [0.1, M]$

2: **for** $l : 1 \to N$ **do**

3: Para cada combinación de media y varianza se computa $V(\nu_i|\mu)$ con $\mu \sim \mathcal{N}(\xi, \sigma^2)$ usando POT y con ello se estima la función de aptitud a minimizar dada en (3.1)

4: Escoger los p_{selec} mejores individuos según su aptitud;

5: Conservar a esos individuos para la siguiente generación y crear hijos a partir de ellos. En este proceso, cada hijo se genera emparejando aleatoriamente a dos padres, y se calcula su media y varianza como el promedio de las medias y varianzas de los progenitores respectivos.

6: Se mutará algún gen, en este caso la media y/o varianza según p_{mut} escogiendo un valor al azar dentro de la grilla.

7: **end for**

Capítulo 4

Cálculo de Baricentros

En los siguientes experimentos numéricos, se va a ilustrar el comportamiento del baricentro débil reverso (RWB) en el caso unidimensional, junto con compararlo contra los otros baricentros y verificar empíricamente las propiedades teóricas que se mencionaron en el capítulo de resultados. Asimismo, se pondrá a prueba el algoritmo diseñado para el baricentro débil (WWB) en el dataset de MNIST, y se comparará solamente con el baricentro de Wasserstein (WB).

4.1. Baricentros en Distribuciones Gaussianas 1D

Este experimento consta de calcular los tres baricentros entre tres distribuciones gaussianas G1,G2,G3 en dos casos diferentes. El primero que se mostrará será el de gaussianas centradas con diferente varianza, tal como se muestra en la Figura 4.1. La tabla 4.1 resume la media y varianza de cada distribución, tanto las G1,G2,G3 como los baricentros. Estos últimos parámetros fueron calculados mediante el algoritmo genético descrito en el Algoritmo 4

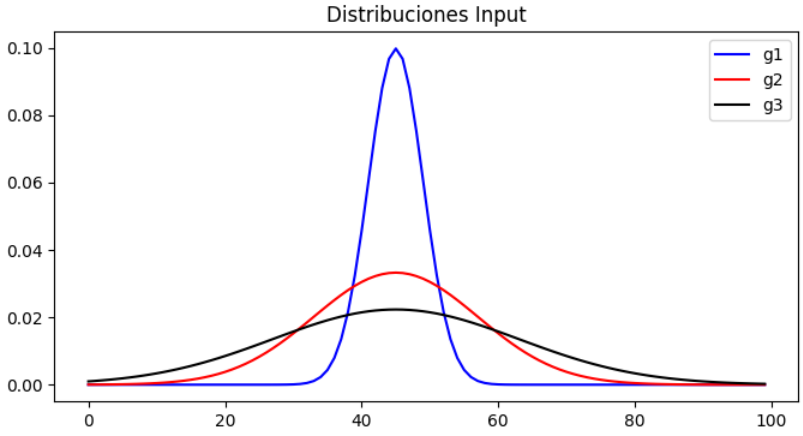


Figura 4.1: Distribuciones input caso centrado

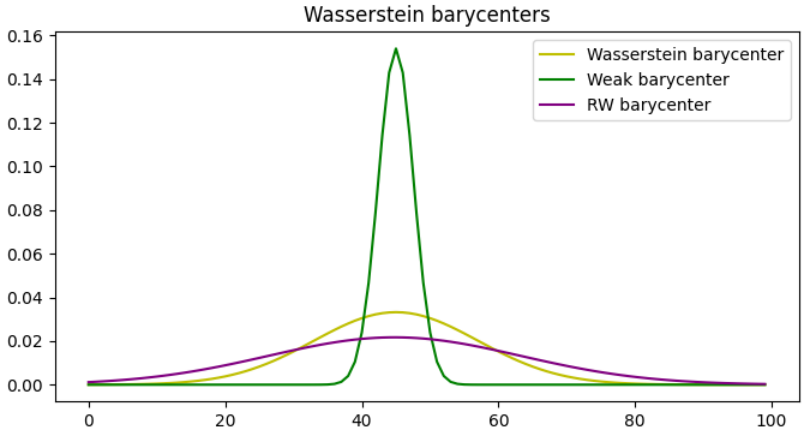


Figura 4.2: Baricentros de Wasserstein entre 3 gaussianas centradas.

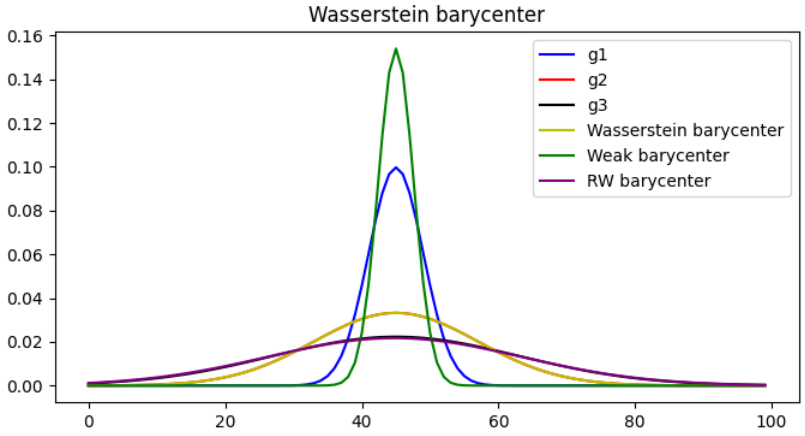


Figura 4.3: Medidas input junto a los 3 baricentros

Tabla 4.1: Media y varianza de los baricentros computados junto a los respectivos parámetros de las distribuciones a promediar caso centrado y M=25.

	WWB	RWB	WB	G1	G2	G3
μ	45	44.78	45	45	45	45
σ^2	2.59	18.54	12	4	12	18

Tabla 4.2: Valores óptimos asociados a cada problema de optimización en el caso centrado y M=25

	$V(G1\|\mu)$	$V(G2\|\mu)$	$V(G3\|\mu)$	Valor Óptimo
Reverso	0.002	0.004	0.007	0.014
Débil	3.310^{-9}	1.710^{-5}	0.016	0.016

Primero que todo cabe destacar que todos los análisis de esta sección relacionados con el orden convexo entre medidas se hará a través de la varianza de cada distribución involucrada,

basado en la observación 2.2.5. Por otra parte, los resultados mostrados anteriormente fueron obtenidos para $M = 25$, lo cual es suficientemente grande en este contexto. El conjunto de soluciones del baricentro débil para el caso gaussiano centrado unidimensional contiene a todas las distribuciones gaussianas cuya varianza es menor o igual a la varianza más pequeña de todas las distribuciones a promediar. Por otra parte, en el caso reverso el conjunto de soluciones contiene a todas las distribuciones gaussianas cuya varianza es mayor o igual a la varianza más grande de todas las distribuciones a promediar, a la vez que es menor o igual a M. En el caso del baricentro débil, se puede ver en la tabla 4.1 que el algoritmo obtuvo una varianza de 2.59, que es menor a todas las varianzas de las medidas input y, por tanto, está por debajo de ellas en orden convexo. Por otro lado, el baricentro débil reverso cumple que su varianza está por sobre todas las varianzas de las distribuciones input y también que es menor o igual a M. En la figura 4.2 se observa una relación de orden convexo entre los tres baricentros, hecho que se ratifica observando las varianzas de estas tres distribuciones en la Tabla 4.1. Específicamente, se ve el orden convexo entre ambos baricentros débiles, tal como se esperaba teóricamente dada la elección de M. Gracias a la figura 4.3 se puede ver que, por un lado, el débil busca concentrar toda la masa en un intervalo que comparten todas las gaussianas. En otras palabras, reúne la información común que comparten las tres gaussianas promediadas. En cambio, el reverso trata de englobar la mayor cantidad de masa posible para contener la mayor cantidad de información concebible de cada medida input. Esto se traduce en que cada medida input tiene información oculta del baricentro, y este trata de resumir toda la información en una sola distribución, tal como lo señala el Teorema 3.2.13. Por último, se observa en la Tabla 4.2 que los problemas de optimización de ambos baricentros débiles valen aproximadamente lo mismo, lo que se esperaba teóricamente por las Observaciones 3.2.9 y 3.2.14.

El segundo caso fue abordar el cálculo de baricentros cuando G1,G2 y G3 no están centradas. En la Tabla 4.3 se muestra, al igual que en el primer caso, los parámetros de cada distribución involucrada. Y en la Figura 4.4 se muestran las distribuciones a promediar.

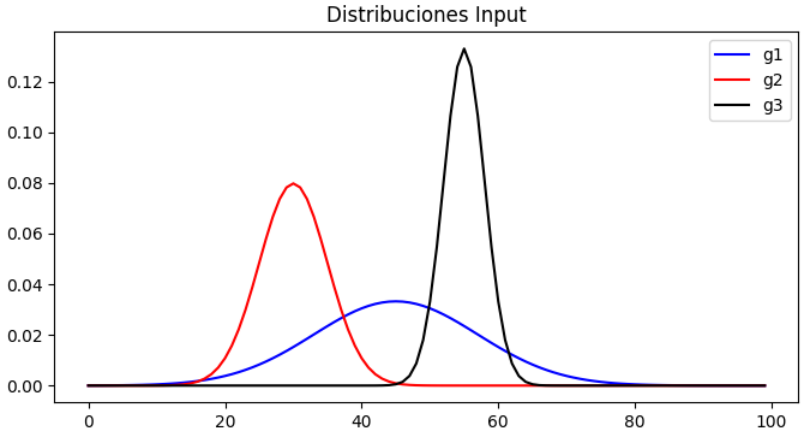


Figura 4.4: Distribuciones input caso no centrado

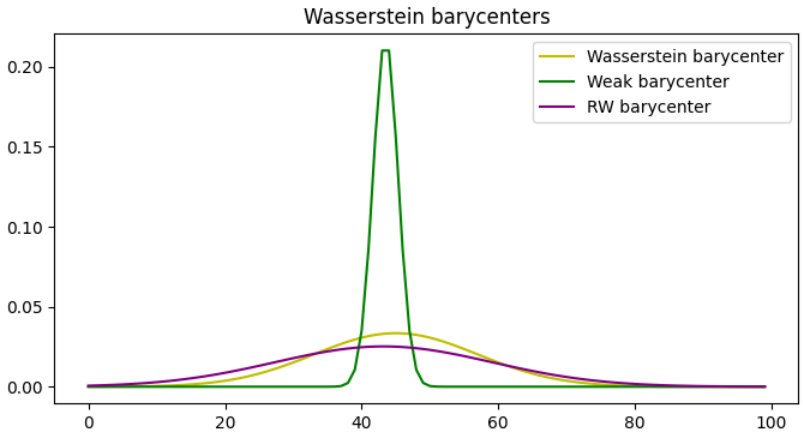


Figura 4.5: Baricentros de Wasserstein entre 3 gaussianas no centradas

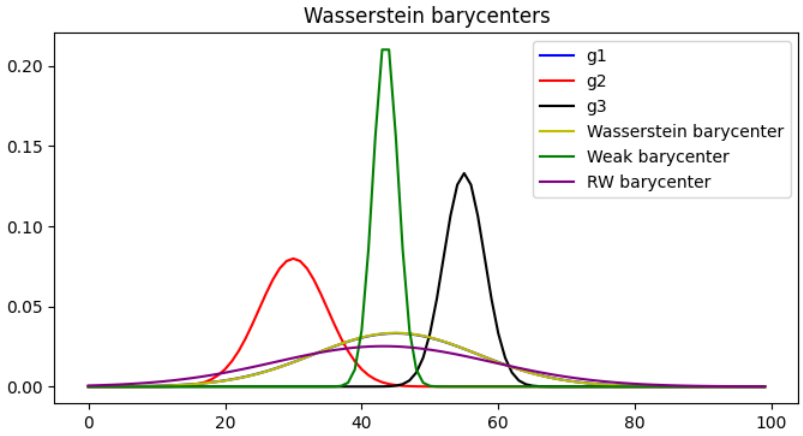


Figura 4.6: Medidas input junto a los 3 baricentros

Tabla 4.3: Media y varianza de los baricentros computados junto a los respectivos parámetros de las distribuciones a promediar caso no centrado.

	WWB	RWB	WB	G1	G2	G3
μ	43.5	43.1974	45	45	30	55
σ^2	1.83	15.8959	11.993	12	5	3

Tabla 4.4: Valores óptimos asociados a cada problema de optimización caso no centrado y M=25

	$V(G1\|\mu)$	$V(G2\|\mu)$	$V(G3\|\mu)$	Valor Óptimo
Reverso	0.91	59.36	45.27	105.56
Débil	0.75	60.74	44.1	105.6

En el caso no centrado, se pueden ver verificaciones empíricas de diferentes proposiciones relacionadas con ambos baricentros. En primer lugar, se tiene que en el caso débil, si se centran todas las variables aleatorias involucradas, viendo la varianza de cada

una en la Tabla 4.3 se concluye que el baricentro débil está por debajo en orden convexo de todas las medidas input centradas, y también cumple la propiedad de la media, pues $\mathbb{E}_\mu X = \frac{1}{3}(\mathbb{E}_{\nu_1} Y_1 + \mathbb{E}_{\nu_2} Y_2 + \mathbb{E}_{\nu_3} Y_3) = (30 + 45 + 55)/3 = 43.3$ y la aproximación da una esperanza de 43.5 según la Tabla 4.3. Por otra parte, es importante notar que el baricentro reverso satisface la propiedad de la media al igual que el baricentro débil, pues se obtuvo una media de 43.19, tal como se muestra en la Tabla 4.3. En segundo lugar el baricentro reverso satisface las proposiciones 3.2.6 y 3.2.7 pues al centrar las medidas y el baricentro reverso, se observa gracias a las varianzas dadas en la Tabla 4.3 que está por sobre todas las medidas input en orden convexo, por ende es solución del problema. Gracias a la Tabla 4.4 se puede observar la igualdad de los problemas de optimización y que se satisface la Observación 3.2.14.

Por la observación 2.2.5 el baricentro reverso no está en relación de orden convexo con dos de las tres medidas input, pues gracias a la tabla 4.3 se puede ver que el RWB y G3,G2 no tienen la misma media. A pesar de lo anterior, como en este caso se satisfacen las hipótesis del Teorema 3.2.13 pues las medidas a promediar son independientes y M es lo suficientemente grande, se concluye que las medidas input siguen siendo variables latentes para el baricentro.

4.1.1. Caso M pequeño

En este experimento se busca mostrar de manera empírica el efecto que tiene sobre las soluciones modificar el valor de M, siendo lo suficientemente pequeño para que no sean válidas ciertas propiedades teóricas del baricentro débil reverso. Para ello se consideran las mismas distribuciones del caso anterior, pero con $M = 6$. En la tabla 4.5 se muestran los valores óptimos de cada problema de optimización. En la figura 4.7 se muestra el baricentro débil reverso, y en la tabla 4.6 se muestran los parámetros de este baricentro junto a los parámetros de las distribuciones a promediar.

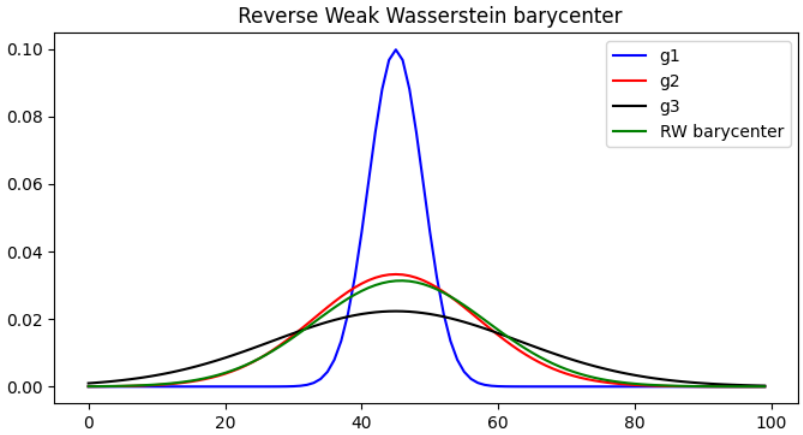


Figura 4.7: Baricentro Reverso para M=6 caso centrado

De acuerdo con las tablas 4.2 y 4.5 se puede ver que el valor del problema de optimización incrementó, por lo que se pierde la igualdad de funciones objetivos entre ambos baricentros débiles, y en consecuencia el Teorema 3.2.13 no aplica en este caso. Los resultados teóricos que permiten obtener alguna relación entre medidas input y el baricentro están dados por los Teoremas 3.2.16 y 3.2.15. Gracias a la Tabla 4.6 se observa que el baricentro reverso está por sobre G1 y G2 en orden convexo, y por ende estas variables son latentes al baricentro. Por

Tabla 4.5: Valores óptimos asociados a cada problema de optimización en el baricentro débil reverso para M=6

	$V(G1\|\mu)$	$V(G2\|\mu)$	$V(G3\|\mu)$	Valor Óptimo
Centrado	0.19	0.188	7.85	8.236

Tabla 4.6: Media y varianza del baricentro débil reverso junto a los respectivos parámetros de las distribuciones a promediar caso centrado y M=6.

	RWB	G1	G2	G3
μ	45.75	45	45	45
σ^2	12.72	4	12	18

la misma tabla se puede concluir que el baricentro está por debajo en orden convexo de G3, por lo tanto, esta medida es variable latente para G3. Este es un ejemplo donde el baricentro débil reverso posee diferente tipo de información de las medidas input de manera simultánea. Es decir, algunas medidas son variables latentes para el baricentro, mientras que para otras medidas resulta ser que el baricentro es variable latente para ellas.

A continuación se exhiben los resultados para el caso no centrado:

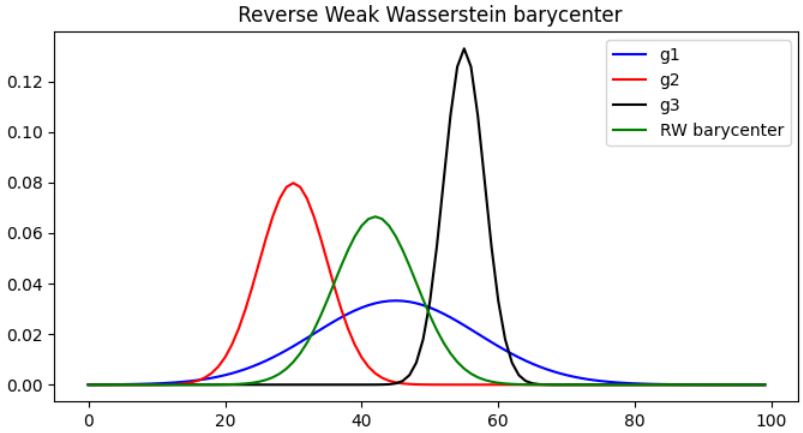


Figura 4.8: Baricentro Reverso para M=6 caso no centrado

Tabla 4.7: Media y varianza del baricentro débil reverso junto a los respectivos parámetros de las distribuciones a promediar caso no centrado y M=6.

	RWB	G1	G2	G3
μ	42	45	30	55
σ^2	6	12	5	3

Considerando el caso no centrado, ninguno de los teoremas desarrollados en la sección de resultados permiten concluir alguna relación de variable latente entre las medidas a promediar y el baricentro. En este caso no hay relación de orden convexo entre las medidas input y el baricentro, pues, gracias a la Tabla 4.7, ninguna esperanza coindice, lo cual es condición

Tabla 4.8: Valores óptimos asociados a cada problema de optimización en el baricentro débil reverso para M=6

	$V(G1\|\mu)$	$V(G2\|\mu)$	$V(G3\|\mu)$	Valor Óptimo
No centrado	15.02	47.99	56.33	119.35

necesaria para que exista orden convexo entre medidas. De acuerdo con la Tabla 4.7 al centrar las medidas involucradas, se observa el mismo comportamiento que en el caso centrado. Esto da a entender que se podría obtener alguna relación como la del Teorema 3.2.13 siempre y cuando las versiones centradas de estas distribuciones cumplan que el baricentro está por sobre las medidas input en orden convexo. Por último, señalar que, por la Tabla 4.8 se observa que los valores de los problemas de optimización incrementaron con respecto a lo visto en la Tabla 4.4, lo cual es esperable porque la elección de M provoca perder propiedades teoricas, como la dada en la Observación 3.2.9.

4.2. Baricentros en MNIST

Se busca computar baricentros en MNIST, los siguientes experimentos consisten en calcularlos entre dos distribuciones correspondientes al 0 y 1. Para ello se empleó el algoritmo de WIN en el caso no débil, mientras que para el caso débil se utilizó el algoritmo desarrollado en 0. En ambos casos se usó para el generador y los potenciales una resnet. Mientras que los transportes fueron parametrizados por una U-net. La configuración de hiperparámetros en cada algoritmo fueron los siguientes:

Baricentro de Wasserstein

- G ITERS, D ITERS, T ITERS = 100, 60, 15
- G LR, D LR, T LR = 3e-4, 3e-4, 3e-4
- BATCH SIZE = 32
- MAX STEPS = 3301

Baricentro Débil

- G ITERS, D ITERS, T ITERS = 230, 50, 10
- G LR, D LR, T LR = 3e-4, 3e-4, 3e-4
- ZC = 1
- Z SIZE = 40
- BATCH SIZE = 32
- MAX STEPS = 2641

En cada iteración del algoritmo, por cada D iters de los poteneciales, se emplean T iters para estimar los mapas estocásticos. Luego se va a la parte de entrenar el generador G por G Iters veces. 1 paso completo del algoritmo toma G ITERS + N*D ITERS, con n la cantidad

de distribuciones a promediar. Se realizan tantos ciclos hasta alcanzar el MAX STEPS. La única diferencia entre las arquitecturas empleadas en ambos algoritmos es que en el caso débil, solo se modifica la capa de entrada de la U-net, añadiendole el parámetro ZC que corresponde a la variable z de los mapas estocásticos. Los parámetros G LR corresponden a los *learning rate* para cada entrenamiento de las redes.

En cuanto al funcionamiento específico del Algoritmo 0 en el caso de imágenes, la distribución no atómica S del algoritmo corresponde a una normal estándar $\mathcal{N}(0, I)$ en el espacio $\mathbb{R}^{16}$. Las distribuciones a promediar son distribuciones empíricas en el espacio $\mathbb{R}^{784}$, que corresponde al espacio donde pertenecen las imágenes de tamaño 1x28x28. En cada iteración del algoritmo, en el primer paso que corresponde a estimar los mapas estocásticos $T_j(x, z)$, se samplea un batch de datos de la distribución S, y se le aplica la función G_ξ, que corresponde al generador. Luego se toma un batch de la distribución empírica y se realizan las optimiciones respectivas para encontrar los parámetros de esta función T_j siguiendo lo indicado en el Algoritmo 0. Para calcular el operador de punto fijo, primero se calculan las proyecciones baricéntricas mediante una estimación de montecarlo en la variable z con Z SIZE muestras la distribución $\mathbb{U}$ del Algoritmo 0 (que en la práctica la variable z se samplea de una distribución normal estándar pero en el espacio de las imágenes) de las funciones $T(x, z)$, y en este caso cada $T(x, z)$ corresponde a una imagen, y el promedio en montecarlo se realiza pixel a pixel. Luego se calcula el operador de punto fijo mediante la ecuación 2.24. En el segundo paso de la optimización, en la regresión para encontrar los parámetros del generador G_ξ se usa el error cuadrático medio entre el output de la red y la estimación de montecarlo del operador de punto fijo mecionada. Cabe destacar que la comparación también es pixel a pixel.

Siguiendo la metodología anterior, sumado a que también se estimó la varianza de cada baricentro de manera empírica. Esto es, samplear 5000 imágenes de las redes generadoras que aproximan cada baricentro, y luego calcular la varianza de la muestra. A continuación se muestran los resultados obtenidos para cada baricentro entre las distribuciones empíricas de 0 y 1 en MNIST.

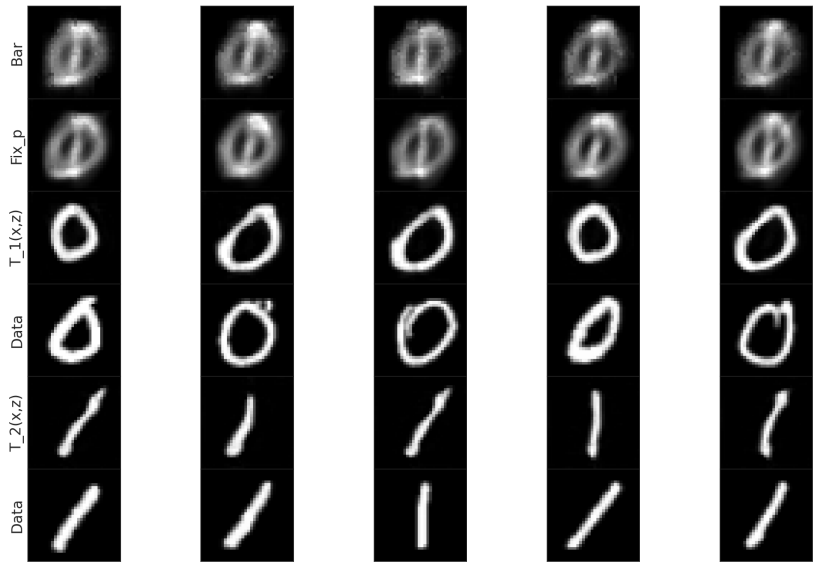


Figura 4.9: Baricentro débil entre el 0 y 1 en MNIST. La primera fila corresponde a sampleos del generador $G_\xi(S)$. La segunda es la función $G(\mu)$, la cuarta y sexta fila corresponden a un dato. El resto de filas corresponde a las funciones $T_j(x, z_i)$ donde para cada $j = 1, 2$ T_j es un mapeo que va desde la imágen dada por el generador al 0 o al 1 segun corresponda.

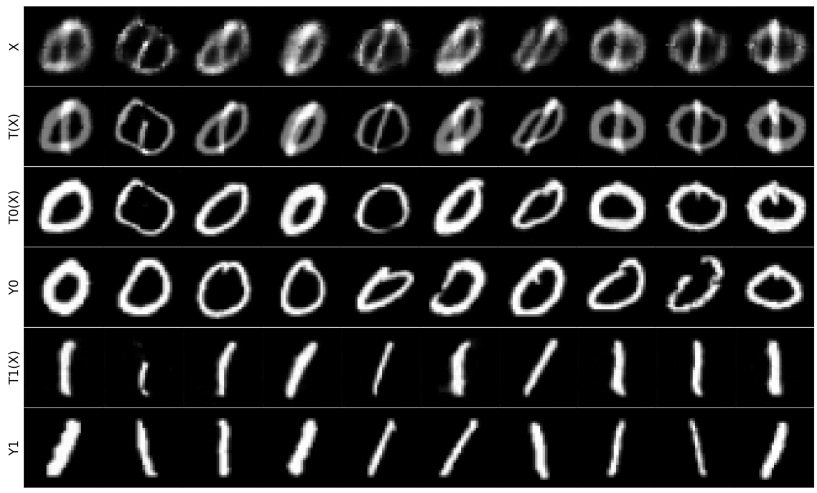


Figura 4.10: Baricentro Entre 0 y 1 en MNIST. La primera fila corresponde a muestras del generador $G_\xi(S)$. La segunda fila representa el operador de punto fijo en este caso. La cuarta y sexta filas corresponden a datos. Las filas restantes corresponden a las funciones $T_j(x)$ donde para cada$j = 1, 2$, T_j es un mapeo que va desde la imagen dada por el generador hasta 0 o 1, según corresponda.

Gracias a las Figuras 4.10 y 4.9 se puede ver como el baricentro de Wasserstein y el débil resumen la información de las distribuciones a promediar. Por una parte, en la figura 4.10 se ve que busca juntar la estructura de ambas imágenes en una sola, promediando los valores de los pixeles. Mientras que en el baricentro débil dado por la Figura 4.9, si bien también busca juntar la estructura de ambos digitos, se contemplan imágenes más estables,

Tabla 4.9: Varianzas de ambos baricentros entre el 0 y 1

	Baricentro Débil	Baricentro de Wasserstein
Varianza	0.0044	0.0822

independientes de la forma de los dígitos a promediar, en otras palabras, los sampleos del baricentro se parecen más entre sí, a diferencia del baricentro de Wasserstein, donde cada imagen sampleada busca ajustarse más a los números que se están promediando. Por ejemplo, en la Figura 4.9 muestra que cada sampleo de este baricentro busca asemejarse más a las imágenes que corresponden a los datos de esa columna. Esto quiere decir que el baricentro débil está más concentrado que el baricentro de Wasserstein, y es más robusto a los outliers de las medidas input. Hecho que se ratifica en la Tabla 4.9 pues la varianza del baricentro débil es menor que la varianza del baricentro de Wasserstein. Además, esta estabilidad se condice con la propiedad de la media que satisface el baricentro débil dada por la Ecuación (2.19).

4.3. Data Augmentation

De manera general lo que se puede hacer para realizar *Data Augmentation* mediante baricentros es considerar diferentes datasets para una misma tarea, por ejemplo clasificación. Luego tomar una clase en particular y obtener las distribuciones de esa clase pero de los diferentes datasets y computar el baricentro de ellas. Para ilustrar esta idea se realizaron 2 experimentos, ambos consisten en considerar la distribución de unos de MNIST y añadirle ruido. La única diferencia entre ellos es el tipo de ruido que se emplea. Posteriormente se calcula el baricentro entre la distribución original y la ruidosa. El baricentro obtenido constituye data sintética.

El primer ruido usado fue tomar una imagen de un uno y cambiar el color de un pixel segun una distribución bernoulli de parámetro 0.1

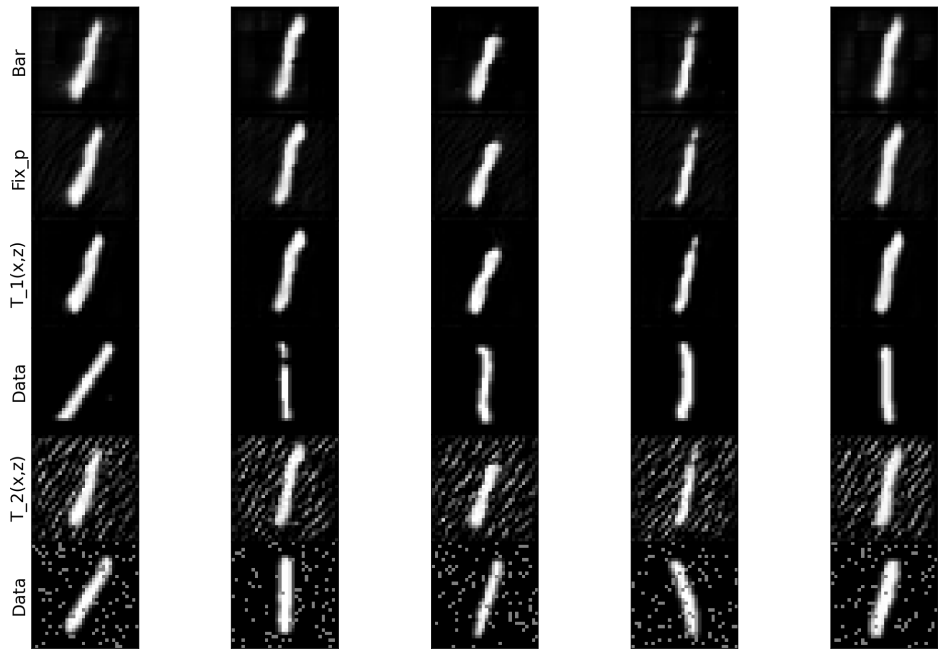


Figura 4.11: Baricentro Débil de Wasserstein entre la distribución de 1 y 1 con ruido.

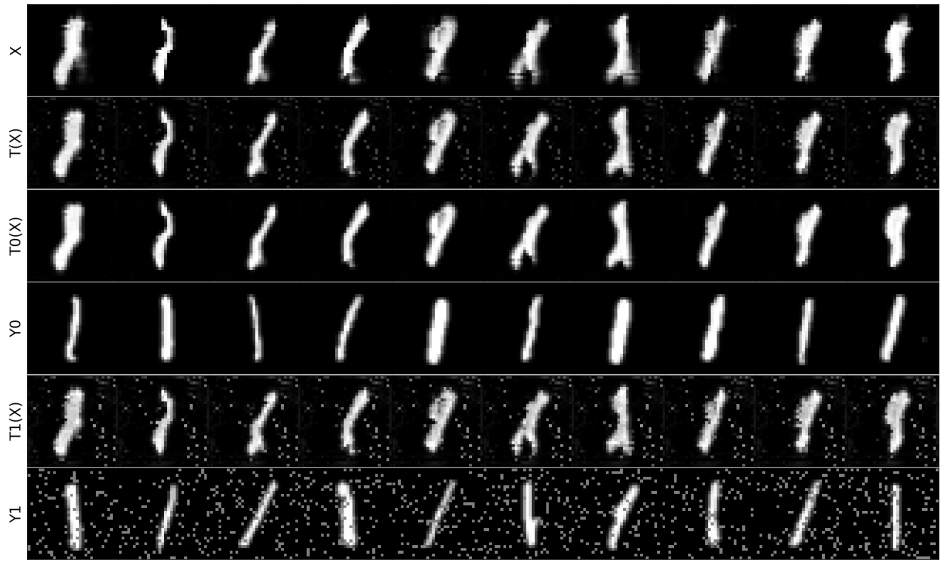


Figura 4.12: Baricentro de Wasserstein entre la distribución de 1 y 1 con ruido.

Tabla 4.10: Varianzas de los baricentros obtenidos en el experimento con el primer ruido.

	Baricentro Débil	Baricentro de Wasserstein
Varianza	0.022	0.084

En las Figuras 4.11 y 4.12 se logra apreciar como se comporta cada baricentro. El de Wasserstein tiende a ser más disperso que el débil, pues en el débil se visualizan unos más similares entre si y la varianza de este baricentro, según la Tabla 4.10 es menor a la del baricentro de Wasserstein.

El segundo ruido usado se construyó tomando una imagen de un uno y cambiar el color de los pixeles blancos a negros según una distribución bernoulli de parámetro 0.45

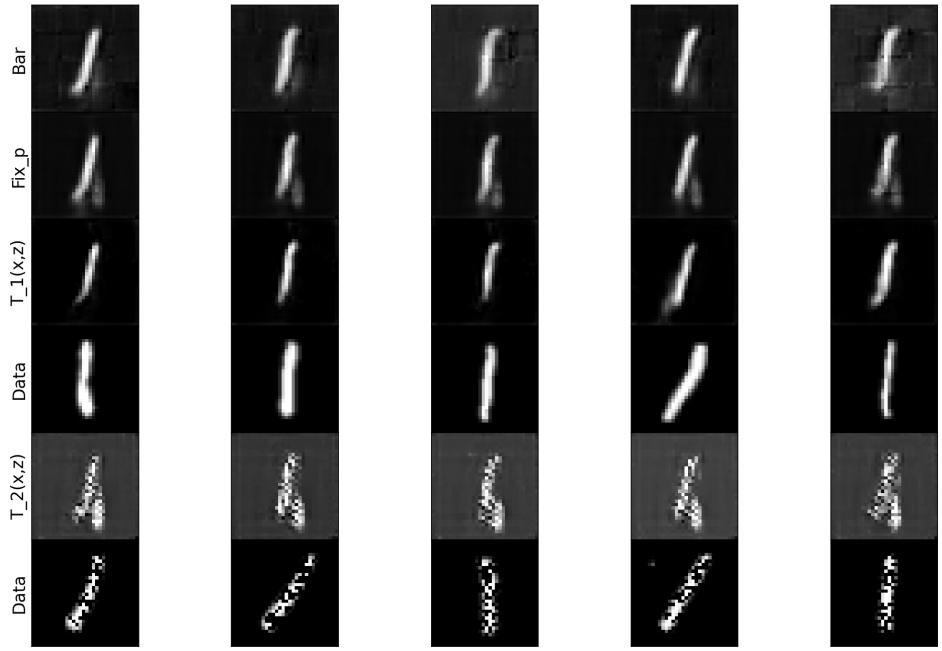


Figura 4.13: Baricentro Débil de Wasserstein entre la distribución de 1 y 1 con segundo ruido empleado.

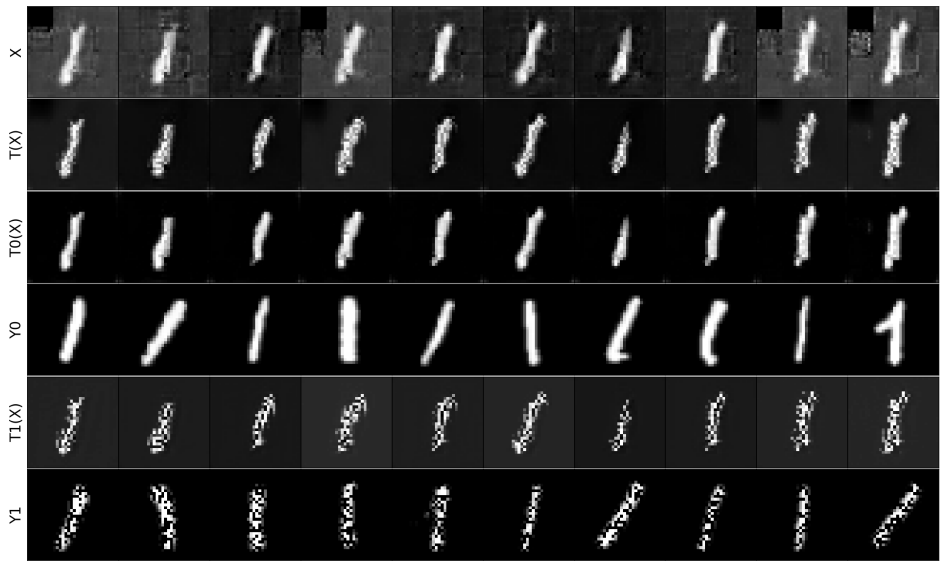


Figura 4.14: Baricentro de Wasserstein entre la distribución de 1 y 1 con segundo ruido empleado.

Tabla 4.11: Varianzas de los baricentros en experimento con ruido 2

	Baricentro Débil	Baricentro de Wasserstein
Varianza	0.015	0.029

En las Figuras 4.13 y 4.14 se logra apreciar cómo se comporta cada baricentro. Existe el mismo análisis que en el primer experimento, solo que en este caso es aún más notoria la dispersión del baricentro de Wasserstein, ya que en la figura 4.14 los samples del baricentro son mucho más ruidosos y dispersos que su contraparte débil. Además, por la Tabla 4.11, se sigue observando que la varianza del baricentro débil es menor que la varianza del baricentro de Wasserstein.

Gracias a ambos experimentos se observa que el baricentro débil logra ser más estable al ruido, por ende más robusto a outliers y genera nueva data que busca capturar cosas en común de las distribuciones que se promediaron, lo cual la hace más útil a la hora de ocupar estos datos para entrenar modelos de *Machine Learning*.

Capítulo 5

Conclusiones

El presente trabajo ha abordado el estudio teórico y algorítmico de los baricentros derivados del transporte óptimo débil, centrándose en potenciales aplicaciones para el aprendizaje de máquinas. Gracias a la asimetría presentada por el problema de transporte óptimo débil, se desarrolló una nueva idea de baricentro, llamada baricentro débil reverso. Se logró satisfactoriamente la demostración de propiedades del baricentro débil reverso, la obtención de un algoritmo eficiente para el cálculo del baricentro débil, así como la evaluación del comportamiento teórico del baricentro de Wasserstein y el baricentro débil en el dataset de MNIST, y finalmente la misma evaluación pero de los tres baricentros en distribuciones gaussianas en una dimensión.

El resultado más destacado revela que todas las medidas a promediar, para M adecuado, se consideran variables latentes para el baricentro débil reverso. Si M es lo suficientemente pequeño, se tiene que todas las medidas que están por debajo en el orden convexo del baricentro débil reverso son variables latentes para el. También se tiene que todas las medidas que están por sobre el baricentro débil reverso en el orden convexo, el baricentro es una variable latente para ellas. Estos hallazgos son cruciales, ya que indica que el baricentro débil reverso recopila información de manera única y diferente al baricentro débil, proporcionando una nueva perspectiva en el contexto del aprendizaje de máquinas. Este hecho se respalda tanto teórica como numéricamente en la tesis. Esto será útil en contextos donde se necesite resumir toda la información proporcionada por las distribuciones de entrada en una sola. A diferencia del baricentro débil, cuya relevancia será evidente en situaciones donde es necesario tener la información común compartida por las medidas de entrada.

A pesar de los logros, la tesis también reconoce una limitación importante: la dificultad para desarrollar un algoritmo eficiente para el cálculo del baricentro débil reverso en más de una dimensión, lo que ha impedido su aplicación directa en el conjunto de datos MNIST. Esta restricción resalta la complejidad inherente del problema y sugiere áreas para investigaciones futuras, las cuales se detallan a continuación.

5.1. Trabajo Futuro

1. Evaluar si existe algún tipo de ecuación que cumpla el baricentro débil reverso, la idea sería obtener algo similar al operador de punto fijo para el caso reverso con el fin de poder obtener un algoritmo.

2. Explorar técnicas de optimización binivel para el cálculo de baricentros.
3. Extender el baricentro reverso de manera similar al weak population baricenter
4. Aplicar el baricentro débil reverso en el contexto de multi source domain adaptation, pues se cree que en este contexto las propiedades de este promedio serán más útiles que el resto.
5. En la misma situación de antes, emplear el algoritmo desarrollado para el caso débil
6. Emplear aquel algoritmo en diferente tipo de datos
7. Adaptar mejor la red generadora G del algoritmo para el baricentro débil con el fin de que funcione mejor en imágenes.
8. ¿Qué pasa con el baricentro débil reverso en presencia de outliers?
9. ¿Es viable pensar en una relación tipo teorema 3.2.13 cuando M es pequeño y cuando las variables aleatorias a promediar no son independientes?.
10. ¿Cómo influye M en las soluciones?, ¿Existirá alguna forma de escoger el M adecuado dependiendo de cuanta información es la que se busca resumir?

Bibliografía

[1] Agueh, M. y Carlier, G., "Barycenters in the wasserstein space," SIAM Journal on Mathematical Analysis, vol. 43, no. 2, pp. 904–924, 2011.

[2] Gozlan, N., Roberto, C., Samson, P.-M., y Tetali, P., "Kantorovich duality for general transport costs and applications," 2015.

[3] Cazelles, E., Tobar, F., y Fontbona, J., "A novel notion of barycenter for probability distributions based on optimal weak mass transport," Advances in Neural Information Processing Systems, vol. 34, pp. 13575–13586, 2021.

[4] Peyré, G., Cuturi, M., *et al.*, "Computational optimal transport: With applications to data science," Foundations and Trends® in Machine Learning, vol. 11, no. 5-6, pp. 355–607, 2019.

[5] Kong, X., Jiang, X., Zhang, B., Yuan, J., y Ge, Z., "Latent variable models in the era of industrial big data: Extension and beyond," Annual Reviews in Control, vol. 54, pp. 167–199, 2022, doi:10.1016/j.arcontrol.2022.09.005.

[6] Courty, N., Flamary, R., Tuia, D., y Rakotomamonjy, A., "Optimal transport for domain adaptation," 2016.

[7] Arjovsky, M., Chintala, S., y Bottou, L., "Wasserstein gan," 2017.

[8] Villani, C. *et al.*, "Optimal transport: Old and new," 2009.

[9] Simon, D. y Aberdam, A., "Barycenters of natural images constrained wasserstein barycenters for image morphing," en Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7910–7919, 2020.

[10] Dognin, P., Melnyk, I., Mroueh, Y., Ross, J., Santos, C. D., y Sercu, T., "Wasserstein barycenter model ensembling," arXiv preprint arXiv:1902.04999, 2019.

[11] Altschuler, J. M. y Boix-Adserà , E., "Wasserstein barycenters are NP-hard to compute," SIAM Journal on Mathematics of Data Science, vol. 4, pp. 179–203, 2022, doi:10.1137/21m1390062.

[12] Fréchet, M., "Les éléments aléatoires de nature quelconque dans un espace distancié," Annales de l'institut Henri Poincaré, vol. 10, no. 4, pp. 215–310, 1948, http://www.numdam.org/item/AIHP_1948___10_4_215_0/.

[13] Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., y Matrán, C., "A fixed-point approach to barycenters in wasserstein space," Journal of Mathematical Analysis and Applications, vol. 441, no. 2, pp. 744–762, 2016.

[14] Backhoff-Veraguas, J. y Pammer, G., "Applications of weak transport theory," Bernoulli, vol. 28, no. 1, pp. 370–394, 2022.

[15] Korotin, A., Selikhanovych, D., y Burnaev, E., "Neural optimal transport," 2023.

[16] Veraguas, J. B., Beiglboeck, M., y Pammer, G., "Existence, duality, and cyclical monotonicity for weak transport costs," 2019.

[17] Strassen, V., "The existence of probability measures with given marginals," The Annals of Mathematical Statistics, vol. 36, no. 2, pp. 423–439, 1965.

[18] Backhoff-Veraguas, J., Beiglböck, M., y Pammer, G., "Weak monotone rearrangement on the line," 2019.

[19] Kallenberg, O. y Kallenberg, O., Foundations of modern probability, vol. 2. Springer, 1997.

[20] Rosenblatt, F., "The perceptron: a probabilistic model for information storage and organization in the brain.," Psychological review, vol. 65, no. 6, p. 386, 1958.

[21] LeCun, Y., Bengio, Y., y Hinton, G., "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.

[22] Samat, N. A., Salleh, M. N. M., y Ali, H., "The comparison of pooling functions in convolutional neural network for sentiment analysis task," en Recent Advances on Soft Computing and Data Mining: Proceedings of the Fourth International Conference on Soft Computing and Data Mining (SCDM 2020), Melaka, Malaysia, January 22-23, 2020, pp. 202–210, Springer, 2020.

[23] He, K., Zhang, X., Ren, S., y Sun, J., "Deep residual learning for image recognition," en Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[24] Ronneberger, O., Fischer, P., y Brox, T., "U-net: Convolutional networks for biomedical image segmentation," en Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241, Springer, 2015.

[25] Bhandari, D., Murthy, C., y Pal, S. K., "Genetic algorithm with elitist model and its convergence," International journal of pattern recognition and artificial intelligence, vol. 10, no. 06, pp. 731–747, 1996.

[26] Kertz, R. P. y Rösler, U., "Complete lattices of probability measures with applications to martingale theory," Lecture Notes-Monograph Series, pp. 153–177, 2000.

[27] Ye, J., Wu, P., Wang, J. Z., y Li, J., "Fast discrete distribution clustering using wasserstein barycenter with sparse support," 2017.

[28] Bigot, J., Cazelles, E., y Papadakis, N., "Data-driven regularization of wasserstein barycenters with an application to multivariate density registration," 2019.

[29] Anderes, E., Borgwardt, S., y Miller, J., "Discrete wasserstein barycenters: Optimal transport for discrete data," 2015.

[30] Cuturi, M. y Peyré , G., "Semidual regularized optimal transport," SIAM Review, vol. 60, pp. 941–965, 2018, doi:10.1137/18m1208654.

[31] Yang, L., Li, J., Sun, D., y Toh, K.-C., "A fast globally linearly convergent algorithm for the computation of wasserstein barycenters," 2020.

[32] Ge, D., Wang, H., Xiong, Z., y Ye, Y., "Interior-point methods strike back: Solving the

wasserstein barycenter problem," 2020.

[33] Cuturi, M. y Doucet, A., "Fast computation of wasserstein barycenters," 2014.

[34] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., y Peyré, G., "Iterative bregman projections for regularized transportation problems," 2014.

[35] Cuturi, M. y Peyré, G., "A smoothed dual approach for variational wasserstein problems," 2015.

[36] Staib, M., Claici, S., Solomon, J., y Jegelka, S., "Parallel streaming wasserstein barycenters," 2017.

[37] Claici, S., Chien, E., y Solomon, J., "Stochastic wasserstein barycenters," 2018.

[38] Dvurechensky, P., Dvinskikh, D., Gasnikov, A., Uribe, C. A., y Nedić, A., "Decentralize and randomize: Faster algorithm for wasserstein barycenters," 2020.

[39] Li, L., Genevay, A., Yurochkin, M., y Solomon, J., "Continuous regularized wasserstein barycenters," 2020.

[40] Fan, J., Taghvaei, A., y Chen, Y., "Scalable computations of wasserstein barycenter via input convex neural networks," 2021.

[41] Korotin, A., Li, L., Solomon, J., y Burnaev, E., "Continuous wasserstein-2 barycenter estimation without minimax optimization," 2021.

[42] Cohen, S., Arbel, M., y Deisenroth, M. P., "Estimating barycenters of measures in high dimensions," 2021.

[43] Chi, J., Yang, Z., Ouyang, J., y Li, X., "Variational wasserstein barycenters with c-cyclical monotonicity," 2022.

[44] Korotin, A., Egiazarian, V., Li, L., y Burnaev, E., "Wasserstein iterative networks for barycenter estimation," arXiv preprint arXiv:2201.12245, 2022.

[45] Ambrosio, L., Gigli, N., y Savaré, G., Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2005.