Transporte óptimo y adaptación de dominio

Un recorrido desde Gaspard Monge hasta el aprendizaje automático

Brian Britos Simmari

Programa de Posgrado de Maestría en Ciencia de Datos y Aprendizaje Automático
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay
Agosto de 2024

Transporte óptimo y adaptación de dominio

Un recorrido desde Gaspard Monge hasta el aprendizaje automático

Brian Britos Simmari

Tesis de Maestría presentada al Programa de Posgrado de Maestría en Ciencia de Datos y Aprendizaje Automático, Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Magíster en Maestría en Ciencia de Datos y Aprendizaje Automático.

Director:
Ph. D. Prof. Mathias Bourel

Director académico:
Ph. D. Prof. Marcelo Fiori

Montevideo – Uruguay
Agosto de 2024

Britos Simmari, Brian

Transporte óptimo y adaptación de dominio / Brian Britos Simmari. - Montevideo: Universidad de la República, Facultad de Ingeniería, 2024.

xi, 125 p. 29, 7cm.

Director:

Mathias Bourel

Director académico:

Marcelo Fiori

Tesis de Maestría – Universidad de la República, Programa de Maestría en Ciencia de Datos y Aprendizaje Automático, 2024.

Referencias bibliográficas: p. 123 – 125.

1. transporte óptimo, 2. adaptación de dominio, 3. problema de Monge-Kantorovich. I. Bourel, Mathias, . II. Universidad de la República, Programa de Posgrado de Maestría en Ciencia de Datos y Aprendizaje Automático. III. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Ph. D. Prof. Badih Ghattas

Ph.D. Prof. Pablo Musé

Ph.D. Prof. José Rafael León

Montevideo – Uruguay
Agosto de 2024

A Sofi, mi compañera de vida.

Agradecimientos

No es sencillo plasmar el esfuerzo requerido para culminar este trabajo, pero es simple reflejar el hecho de que no lo podría haber hecho solo, es por eso que a continuación quiero agradecer a varias personas que fueron fundamentales para la conclusión de esta tesis.

En primer lugar, quiero expresar mi más profundo agradecimiento a Sofi, quien ha sido, es y siempre será un pilar fundamental y un apoyo invaluable en todo este proceso. Ella me ha acompañado desde el inicio de esta "locura"de estudiar matemática y conoce todos los altibajos que he atravesado. Sofi ha sido y seguirá siendo mi refugio, el lugar donde siempre encuentro el impulso necesario para seguir adelante.

En segundo lugar, quiero expresar mi agradecimiento a mi tutor, Mathias Bourel, quien a pesar de no conocerme me ofreció la oportunidad de trabajar juntos, lo cual resultó ser una experiencia sumamente gratificante. Espero que podamos seguir colaborando en el futuro.

Agradezco también a mi director académico Marcelo Fiori por guiarme a lo largo de la maestría a elegir el mejor conjunto de cursos, así como simplificarme mucho la burocracia de la misma.

No puedo dejar de agradeces a mis padres, Marcelo y Liliam, por su apoyo desde hace 28 años. Desde mi infancia hasta el presente, han estado a mi lado en cada paso del camino.

También quiero agradecer a mis amigos, Tony, Martín, Cynthia, Anto y Santiago, así como a mis compañeros de Scanntech: Nachito, Diego, Hernan y Ramaro, quienes me apoyaron a lo largo del camino.

Finalmente quiero agradecer a Badih, Pablo y Chichi por aceptar ser el tribunal de esta tesis.

Cree que puedes y ya estarás a mitad de camino.

Theodore Roosevelt

RESUMEN

En esta tesis abordamos el problema del transporte óptimo y su aplicación en la adaptación de dominio, presentando un enfoque integral que abarca tanto los fundamentos matemáticos como diversas aplicaciones prácticas.

El transporte óptimo es un área de la matemática que busca minimizar el costo asociado con mover una distribución de masa desde una posición inicial hasta una posición destino. Este costo puede estar basado en diferentes métricas, siendo la distancia euclídea uno de los ejemplos más comunes. Este problema fue formulado inicialmente por el matemático francés Gaspard Monge en el siglo XVIII y más tarde reformulado y extendido por el matemático ruso Leonid Kantorovich en el siglo XX.

Una aplicación reciente en el aprendizaje automático es en el problema de la adaptación de dominio. Este problema consiste en aplicar un modelo de aprendizaje automático entrenado en un dominio fuente, con amplia disponibilidad de datos etiquetados, a un dominio objetivo posiblemente distinto donde los datos etiquetados son escasos o inexistentes. Exploramos cómo utilizar el transporte óptimo para abordar el problema de la adaptación de dominio.

Además de ser una recopilación bibliográfica sobre estos dos temas, proponemos un procedimiento para abordar el problema de la adaptación de dominio cuando el modelo es una regresión lineal simple y los dominios difieren a través de una rotación, donde realizamos varias simulaciones para ponerlo a prueba. Finalmente ponemos en práctica los conocimientos adquiridos a través de experimentos mostrando cómo utilizar el transporte óptimo en la adaptación de dominio sobre conjuntos de datos reales: transferencia de color entre dos imágenes y adaptación de dominio de un clasificador sobre los conjuntos de dígitos MNIST y USPS.

Palabras claves:
transporte óptimo, adaptación de dominio, problema de Monge-Kantorovich.

ABSTRACT

In this thesis, we address the problem of optimal transport and its application in domain adaptation, presenting a comprehensive approach that encompasses both mathematical foundations and various practical applications.

Optimal transport is a mathematical area that seeks to minimize the cost associated with moving a mass distribution from an initial position to a destination position. This cost can be based on different metrics, with Euclidean distance being one of the most common examples. This problem was initially formulated by the French mathematician Gaspard Monge in the 18th century and later reformulated and extended by the Russian mathematician Leonid Kantorovich in the 20th century.

One of the most recent applications of optimal transport in machine learning is in the problem of domain adaptation. This problem involves applying a machine learning model trained on a source domain, with ample availability of labeled data, to a target domain that may be different and where labeled data is scarce or non-existent. We explore how to use optimal transport to address the problem of domain adaptation.

In addition to being a bibliographic compilation on these two topics, we propose a procedure to tackle the problem of domain adaptation when the model is a simple linear regression and the domains differ through a rotation. We conduct several simulations to test it. Finally, we put the acquired knowledge into practice through experiments, demonstrating how to use optimal transport in domain adaptation on real datasets: color transfer between two images and domain adaptation of a classifier on the MNIST and USPS digit datasets.

Keywords:
optimal transport, domain adaptation, Monge-Kantorovich problem.

Tabla de contenidos

Introducción

El transporte óptimo es un campo de la matemática que tiene implicaciones profundas y extensas en una variedad de dominios, incluyendo la economía, la física, y, más recientemente, el aprendizaje automático. La teoría del transporte óptimo se ocupa de encontrar la forma más eficiente de mover recursos de un estado a otro, bajo ciertas restricciones y costos. Originalmente fue formulado por Gaspard Monge en el siglo XVIII (Monge, 1781) y luego refinada por Leonid Kantorovich en el siglo XX (Kantorovich, 1942). La teoría ha evolucionado significativamente y ha encontrado nuevas aplicaciones en la actualidad.

Una de las aplicaciones más prometedoras y desafiantes del transporte óptimo en el aprendizaje automático es la adaptación de dominio. La adaptación de dominio se refiere al problema de aplicar un modelo de aprendizaje automático entrenado en un dominio fuente, con amplia disponibilidad de datos etiquetados, a un dominio objetivo, posiblemente distinto. Estas diferencias suelen ser provocadas por un cambio en la distribución subyacente a los datos entre el momento de entrenar un modelo y el momento de probarlo. Este desafío es particularmente relevante en situaciones donde la recolección de datos etiquetados es costosa, difícil, o impracticable.

La motivación detrás de la integración del transporte óptimo en la adaptación de dominio surge de la necesidad de alinear las distribuciones de los datos en los dominios fuente y objetivo de manera eficiente. Al minimizar el costo de transportar una distribución de probabilidad a otra, el transporte óptimo ofrece un marco matemático riguroso para cuantificar y minimizar las diferencias entre los dominios. Esto facilita la transferencia de conocimiento del dominio fuente al dominio objetivo, mejorando así el rendimiento del modelo de aprendizaje automático en el dominio objetivo, incluso en ausencia de una gran cantidad de datos etiquetados.

Esta tesis explora el potencial del transporte óptimo como una herramienta para abordar el desafío de la adaptación de dominio en el aprendizaje automáti-

co. Al examinar tanto los fundamentos teóricos del transporte óptimo como sus aplicaciones prácticas en la adaptación de dominio, este trabajo busca ser un completo comienzo para quienes quieren adentrarse en este campo.

En el capitulo recopilamos los preliminares matemáticos básicos y notaciones que serán de utilidad en los siguientes capítulos.

En el capítulo 2 se presentan los fundamentos teóricos del problema de transporte óptimo en la formulación de Monge y en la formulación de Kantorovich. Se mostrará parte del contexto histórico que motivo dichos estudios. Luego se introducirán el problema de Kantorovich dual y el transporte óptimo regularizado por entropía.

En el capitulo 3 introduciremos la parte algorítmica, así como las técnicas necesarias para usar el transporte óptimo en la práctica. Se proporcionará una base sólida para comprender cómo se aplica esta teoría en el contexto actual.

El capítulo 4 se centra en la adaptación de dominio, comenzando con su definición y taxonomía. Estudiaremos este problema en los contextos supervisado y no supervisado, analizando un problema de regresión y de clasificación respectivamente.

Finalmente, en el capítulo 5 se presentarán experimentos con datos reales que muestren la utilidad del transporte óptimo y de la adaptación de dominio. Comenzaremos con una aplicación sencilla en regresión lineal simple para entender la geometría del transporte óptimo, proponiendo un procedimiento para abordar el problema de adaptación de dominio cuando los dominios difieren en una rotación. Comprobamos mediante una simulación exhaustiva que el procedimiento propuesto tiene un buen desempeño en diferentes condiciones. Luego, aplicaremos el transporte óptimo entre dos conjuntos de datos reales: transferencia de color entre dos imágenes y adaptación de dominio de un clasificador sobre los conjuntos de dígitos MNIST y USPS.

Capítulo 1

Preliminares matemáticos

En este capítulo se introducirán los preliminares matemáticos y notaciones necesarios para leer y comprender esta tesis. Se asumirá que el lector posee un conocimiento básico en teoría de conjuntos, cálculo, topología y probabilidad. Estos conceptos servirán como base teórica sobre la cual se construirán los argumentos y desarrollos presentados a lo largo de la tesis.

1.1. Medida y probabilidad

Supondremos que $(\mathcal{X}, d)$ es un espacio métrico compacto y denotaremos $\mathcal{C}^0(\mathcal{X})$ al espacio de funciones continuas de $\mathcal{X}$ a $\mathbb{R}$, es decir,

$$\mathcal{C}^0(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \ \mid \ f \text{ es continua}\},$$

donde consideramos la distancia d en $\mathcal{X}$ y la distancia usual en $\mathbb{R}$.

En reiteradas ocasiones trabajaremos con funciones continuas no negativas, por lo que introducimos la siguiente notación:

$$\mathcal{C}^0_+(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \ \mid \ f \text{ es continua y } f \geq 0\}.$$

Una σ-álgebra es un subconjunto $\Sigma \subseteq \mathscr{P}(\mathcal{X})$, el conjunto de partes de $\mathcal{X}$, que cumple las siguientes propiedades:

- $\mathcal{X} \in \Sigma$.
- Si $A \in \Sigma$ entonces su complemento también pertenece a Σ, es decir, $A^c \in \Sigma$.

- Si $\{A_k\}_k$ es una familia numerable de conjuntos pertenecientes a Σ entonces $A = \bigcup_k A_k$ también pertenece a Σ.

Dada una σ-álgebra Σ en $\mathcal{X}$, una función $\mu : \Sigma \to \mathbb{R}$ es una medida no negativa si cumple las siguientes propiedades:

- $\mu(\emptyset) = 0$
- Para todo $A \in \Sigma$ se tiene que $\mu(A) \geq 0$
- Si $\{A_k\}_{k=1}^{\infty}$ es una familia numerable de conjuntos disjuntos dos a dos, entonces
$$\mu\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \sum_{k=0}^{+\infty} \mu(A_k).$$

Diremos que una medida μ es finita si $\mu(\mathcal{X}) < \infty$, es decir, la medida de todo el espacio es finita. A continuación definiremos una clase particular de medidas, denominadas medidas de Radon.

Definición 1.1 (Medida de Radon). *Una medida no negativa μ es una medida de Radon si:*

1. *Para todo abierto $U \in \mathcal{X}$, se tiene que*
$$\mu(U) = \sup\{\mu(K) \ \ | K \subseteq U \text{ y } K \text{ es compacto.}\}$$
2. *Para todo punto $x \in \mathcal{X}$ existe un entorno U_x tal que $\mu(U_x) < \infty$.*

Ejemplos de medidas de Radon pueden ser la medida de Lebesgue en $\mathbb{R}^n$, la medida de Haar en grupos topológicos compactos o las medidas de probabilidad, entre otras.

En este trabajo consideramos únicamente medidas de Radon no negativas. Además, por simplicidad, cuando digamos 'medida' nos estaremos refiriendo a medidas de Radon no negativas.

Al espacio de medidas de Radon finitas sobre $\mathcal{X}$ lo denotaremos $\mathcal{M}^+(\mathcal{X})$. Observar que en el caso donde $\mathcal{X}$ es un espacio (métrico) compacto tenemos que toda medida de Radon es finita ya que si μ es una medida de Radon entonces para todo $x \in \mathcal{X}$ existe un entorno U_x tal que $\mu(U_x) < \infty$. Luego $\{U_x\}_{x\in\mathcal{X}}$ es un cubrimiento por abiertos de $\mathcal{X}$. Como este es compacto, existe un subcubrimiento finito $\{U_{x_k}\}_{k=1}^{K}$. Finalmente $\mu(\mathcal{X}) \leq \mu\Big(\bigcup_{k=1}^{K} U_{x_k}\Big) \leq \sum_{k=1}^{K} \mu(U_{x_k}) < \infty$.

Una clase particular de medidas de Radon finitas son las medidas de probabilidad:

Definición 1.2 (Medida de probabilidad). *Diremos que una medida de Radon finita sobre $\mathcal{X}$ es una medida de probabilidad si $\mu(\mathcal{X}) = 1$.*

Al espacio de medidas de probabilidad lo denotaremos

$$\mathcal{P}(\mathcal{X}) = \{\mu \in \mathcal{M}^+(\mathcal{X}) \ \mid \ \mu(\mathcal{X}) = 1\}.$$

Es claro que $\mathcal{P}(\mathcal{X}) \subset \mathcal{M}^+(\mathcal{X})$ y que la otra inclusión es estricta. De todos modos para cada medida de Radon finita podemos construir una medida de probabilidad normalizando. Explícitamente: si μ es una medida de Radon finita en $\mathcal{X}$ podemos definir la medida $\hat{\mu} = \frac{1}{\mu(\mathcal{X})}\mu$, la cual cumple que $\hat{\mu}(\mathcal{X}) = 1$, es decir, $\hat{\mu}$ es una medida de probabilidad.

Un concepto clave en teoría de la media es el de medidas absolutamente continuas:

Definición 1.3. *Sean μ y ν dos medidas en un espacio métrico compacto $\mathcal{X}$. Se dice que μ es absolutamente continua respecto a ν si para todo boreliano B tal que $\nu(B) = 0$ se tiene que $\mu(B) = 0$. Cuando μ sea absolutamente continua respecto a ν lo denotaremos $\mu \ll \nu$.*

El concepto anterior es importante en el teorema siguiente:

Teorema 1.4 (Derivada de Radon-Nikodym). *Sean μ y ν dos medidas en un espacio métrico compacto $\mathcal{X}$. Si $\mu \ll \nu$ entonces existe una única función medible $f : \mathcal{X} \to [0, \infty)$ tal que*

$$\nu(B) = \int_B f d\mu \quad \textit{para todo boreliano } B.$$

La función f se llama derivada de Radon-Nikodym de ν respecto a μ y denotaremos $f = \frac{d\nu}{d\mu}$.

En probabilidad, es muy frecuente utilizar la derivada de Radon-Nikodym: sean $\mathcal{X} \subset \mathbb{R}$ un conjunto compacto y $P \in \mathcal{P}(\mathcal{X})$ una medida de probabilidad que es absolutamente continua respecto a la medida de Lebesgue m. Si X es una variable aleatoria con distribución P, entonces la derivada de Radon-Nikodym

respecto a la medida de Lebesgue es lo que solemos llamar "densidad", es decir, $\frac{dP}{dm} = f_X$. Luego, la medida inducida por X bajo P, denotada P_X es:

$$P_X(A) = P(X \in A) = \int_A f_X(x)dx \quad \text{para todo boreliano } A.$$

Cuando trabajemos con medidas en $\mathbb{R}$ sera de utilidad la función acumulativa, definida como

Definición 1.5 (Función acumulativa). *Dada una medida $\mu \in \mathcal{M}^+(\mathbb{R})$, su función acumulativa es una función $\mathcal{C}_\mu : \mathbb{R} \to [0,1]$ definida como*

$$\mathcal{C}_\mu(x) = \int_{-\infty}^{x} d\mu.$$

Su pseudo-inversa $\mathcal{C}_\mu^{-1} : [0,1] \to \mathbb{R} \cup \{-\infty\}$ se define por

$$\mathcal{C}_\mu^{-1}(r) = \min_x \Big\{ x \in \mathbb{R} \cup \{-\infty\} \quad | \quad \mathcal{C}_\mu(x) \geq r \Big\}.$$

Observación 1.6. *La función acumulativa suele también llamarse función de distribución mientras que su pseudo-inversa suele llamarse función de cuantil.*

A continuación mostraremos un resultado que necesitaremos en el capítulo 2 para encontrar una forma cerrada de la distancia de Wasserstein entre dos medidas de probabilidad en $\mathbb{R}$.

Proposición 1.7. *Si μ es una medida de probabilidad en $\mathbb{R}$ y $\mathcal{U}$ es la distribución uniforme en $[0,1]$ entonces*

$$(\mathcal{C}_\mu)_\sharp^{-1}\mathcal{U} = \mu,$$

y por lo tanto

$$(\mathcal{C}_\mu)_\sharp\mu = \mathcal{U}.$$

Demostración. Como μ es una medida positiva entonces $\mathcal{C}_\mu$ es una función creciente. Si denotemos $\gamma = (\mathcal{C}_\mu)^{-1}\mathcal{U}$, queremos probar que $\gamma = \mu$, lo cual es

equivalente a $\mathcal{C}_\gamma = \mathcal{C}_\mu$. Tenemos

$$\begin{aligned}
\mathcal{C}_\gamma(x) &= \int_{-\infty}^{x} d\gamma \\
&= \int_R \mathbf{1}_{(-\infty,x]} d\Big((\mathcal{C}_\mu^{-1})_\sharp \mathcal{U}\Big) \\
&= \int_0^1 \mathbf{1}_{(-\infty,x]}\Big((\mathcal{C}_\mu^{-1}(z))\Big) dz \\
&= \int_0^1 \mathbf{1}_{[0,\mathcal{C}_\mu(z)]} dz \\
&= \mathcal{C}_\mu(x).
\end{aligned}$$

□

El resultado de la proposición 1.7 es frecuentemente utilizado para sortear datos de una distribución deseada. Por ejemplo, supongamos que queremos sortear datos de una distribución exponencial con densidad $p(x) = \lambda e^{-\lambda x}$. Primero calculamos la distribución acumulativa, que en este caso es $C_\mu(x) = 1 - e^{-\lambda x}$. Luego tenemos que encontrar la inversa de la función acumulativa, que, para la distribución exponencial, tiene una forma cerrada: $C_\mu^{-1}(u) = -\frac{1}{\lambda}\ln(1-u)$. Luego, sorteamos datos aleatorias $\{u_1, \ldots, u_n\}$ de la distribución uniforme $\mathcal{U}([0,1])$. Finalmente transformamos los datos sorteados con la inversa de la función acumulativa, obteniendo así un conjunto $\{x_1, \ldots, x_n\}$ donde $x_i = C_\mu^{-1}(u_i) = -\frac{1}{\lambda}\ln(1-u_i)$ para todo $i = 1, \ldots, n$. Por la proposición 1.7 podemos asegurar que la muestra $\{x_i, \ldots, x_n\}$ es un sorteo de la distribución exponencial de distribución $p(x) = \lambda e^{-\lambda x}$.

Introducimos a continuación la función indicatriz, que nos será de utilidad al trabajar con medidas discretas:

Definición 1.8. *Dado un conjunto $A \subset \mathcal{X}$ de un espacio métrico compacto, la función indicatriz se define como*

$$\boldsymbol{1}_A(x) = \begin{cases} 1 & si \quad x \in A, \\ 0 & si \quad x \notin A. \end{cases}$$

Consideraremos reiteradamente el caso donde $A = \{x\}$ consiste en un único punto. En tal caso lo notaremos como delta de Dirac δ_x.

Las funciones de interés en teoría de la medida son las llamadas funciones medibles, cuya definición es:

Definición 1.9. *Dado un espacio métrico compacto $(\mathcal{X}, d)$ y una medida $\mu \in \mathcal{M}^+(\mathcal{X})$ diremos que una función $f : \mathcal{X} \to \mathbb{R}$ es medible si $f^{-1}(A) \subset \mathcal{X}$ es medible para todo A medible de $\mathbb{R}$.*

En particular, las funciones $f : \mathcal{X} \to \mathbb{R}$ continuas son medibles. En teoría de la medida, cuando decimos que dos funciones son iguales, en realidad nos referimos a que son iguales salvo en un conjunto de medida nula.

Definición 1.10. *Sea $(\mathcal{X}, d)$ un espacio métrico compacto y μ una medida en $\mathcal{X}$. Diremos que dos funciones medibles $f, g : \mathcal{X} \to \mathbb{R}$ son iguales en casi todo punto si el conjunto de puntos donde son distintas tiene medida (según μ) nula. En tal caso escribiremos $f = g \ \ \mu-ctp.$*

Definimos que dos medidas μ y ν sean iguales $\mu-$ctp cambiando f y g por μ y ν en la definición anterior.

En reiteradas ocasiones integraremos funciones continuas $f \in \mathcal{C}^0(\mathcal{X})$ contra medidas $\mu \in \mathcal{M}^+(\mathcal{X})$. Para simplificar la lectura consideramos la siguiente notación:

$$\langle \varphi, \mu \rangle = \int_{\mathcal{X}} \varphi d\mu.$$

A los espacios $\mathcal{M}^+(\mathcal{X})$ y $\mathcal{P}(\mathcal{X})$ los dotamos con la topología débil, para la cual necesitamos el concepto de convergencia débil.

Definición 1.11 (Convergencia débil). *Sea $\mathcal{X}$ un espacio métrico compacto. Diremos que una sucesión de medidas de Radon $\{\mu_k\}_{k\in\mathbb{N}} \subset \mathcal{M}^+(\mathcal{X})$ converge débilmente a $\mu \in \mathcal{M}^+(\mathcal{X})$ si para toda función continua $\varphi \in \mathcal{C}^0(\mathcal{X})$ se cumple que*

$$\int_{\mathcal{X}} \varphi d\mu_k \xrightarrow{k} \int_{\mathcal{X}} \varphi d\mu,$$

o equivalentemente

$$\langle \varphi, \mu_k \rangle \xrightarrow{k} \langle \varphi, \mu \rangle.$$

En tal caso escribiremos $\mu_k \rightharpoonup \mu$.

Como $\mathcal{P}(\mathcal{X}) \subset \mathcal{M}^+(\mathcal{X})$, la definición anterior incluye a las medidas de probabilidad.

Fijados μ una medida en $\mathcal{M}^+(\mathcal{X})$, $h_1, \ldots, h_n$ funciones en $\mathcal{C}^0(\mathcal{X})$ y $\epsilon > 0$ definimos los siguientes conjuntos:

$$U_{\epsilon,h_1,\ldots,h_n}(\mu) = \Big\{\nu \in \mathcal{M}^+(\mathcal{X}) \;\; | \;\; \Big|\int_{\mathcal{X}} h_i d\nu - \int_{\mathcal{X}} h_i d\mu\Big| \leq \epsilon, i = 1, \ldots, n\Big\}.$$

Tenemos una definición análoga para cuando $\mu \in \mathcal{P}(\mathcal{X})$, cambiando $\nu \in \mathcal{M}^+(\mathcal{X})$ por $\nu \in \mathcal{P}(\mathcal{X})$. Se puede probar que la familia de conjuntos $\{U_{\epsilon,h_1,\ldots,h_n}\}(\mu)$ forma una base de entornos de $\mathcal{M}^+(\mathcal{X})$. Esto conlleva a la siguiente definición.

Definición 1.12 (Topología débil). *La topología débil en $\mathcal{M}^+(\mathcal{X})$ y en $\mathcal{P}(\mathcal{X})$ es la topología inducida por la base de entornos $U_{\epsilon,h_1,\ldots,h_n}(\mu)$.*

Con frecuencia nos encontramos en la siguiente situación: si $\mathcal{X}$ e $\mathcal{Y}$ son dos espacios métricos compactos, $T : \mathcal{X} \to \mathcal{Y}$ es una función medible y μ es una medida en $\mathcal{X}$. ¿Existe una medida $\nu \in \mathcal{M}^+(\mathcal{Y})$ inducida por T tal que $\nu(B) = \mu(T^{-1}(B)$ para todo B boreliano en $\mathcal{Y}$?. La respuesta es afirmativa y motiva la siguiente definición.

Definición 1.13 (Push-Forward). *Sean $\mathcal{X}$ e $\mathcal{Y}$ espacios métricos compactos, $\mu \in \mathcal{M}^+(\mathcal{X})$ y $T : \mathcal{X} \to \mathcal{Y}$ una función medible. El push-forward de μ dado por T es la medida $T_\sharp\mu \in \mathcal{M}^+(\mathcal{Y})$ definida por*

$$\int_{\mathcal{Y}} \varphi dT_\sharp\mu = \int_{\mathcal{X}} \varphi \circ T d\mu \quad \forall \varphi \in \mathcal{C}^0(\mathcal{Y})$$

Podemos escribir la igualdad anterior de forma más compacta como $\langle \varphi, T_\sharp\mu\rangle = \langle \varphi \circ T, \mu\rangle$.

Observación 1.14. *Equivalentemente, la medida $T_\sharp\mu$ se puede definir como la medida que cumple que para todo Boreliano $B \subset \mathcal{Y}$ se tiene que*

$$T_\sharp\mu(B) = \mu(T^{-1}(B)).$$

La figura 1.1 bosqueja la situación.

Ejemplo 1.15. *Si $\mathcal{Y} = \{y_1, \ldots, y_n\}$ entonces $T_\sharp\mu = \sum_{1\leq k\leq n} \mu(T^{-1}\{y_k\})\delta_{y_k}$.*

Ejemplo 1.16. *Una aplicación usual de la medida push-forward es la de definir una medida "natural" en un espacio. Por ejemplo, si $\mathcal{X} = [0, 2\pi) \subset \mathbb{R}$ e $\mathcal{Y} =$*

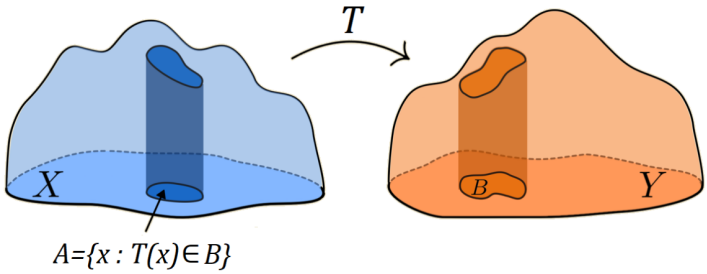


Figura 1.1: Bosquejo del operador push-forward. En azul la distribución μ, en naranja la distribución $\nu = T_\sharp\mu$. Se muestra un conjunto $B \subset \mathcal{Y}$ y el correspondiente $A = T^{-1}(B)$. Imagen tomada del trabajo de Matthew Thorpe (Thorpe, 2018).

$\mathbb{S}^1 := \{(y_1, y_2) \in \mathbb{R}^2 \mid y_1^2 + y_2^2 = 1\}$ *es el círculo unitario en el plano, entonces podemos "llevar" la medida de Lebesgue* m *de* $\mathbb{R}$ *restricta a* $[0, 2\pi)$ *y obtener una medida* λ *en* $\mathbb{S}^1$*. La función medible a utilizar es* $T(x) = (cos(x), sen(x))$*, que es medible por ser continua. Por ejemplo, si en* $\mathbb{S}^1$ *consideramos el arco de circunferencia* s_1 *que comienza en* $(1, 0)$ *y termina en* $(0, 1)$ *entonces sabemos que su longitud es* $\frac{\pi}{2}$ *y por lo tanto*

$$\lambda(s_1) = m\Big(T^{-1}(s_1)\Big) = m\Big(\Big[0, \frac{\pi}{2}\Big]\Big) = \frac{\pi}{2}.$$

La figura 1.2 es un bosquejo de la situación. Con esta construcción la longitud total de S^1 *sería* 2π*, pero podemos recuperar la longitud usual definiendo* $\bar{\lambda} = \frac{1}{2\pi}\lambda$*, así tenemos que* $\bar{\lambda}(\mathbb{S}^1) = 1$*.*

Observar que si quisiéramos restringir la medida de Lebesgue de $\mathbb{R}^2$ *a* $\mathbb{S}^1$*, todo conjunto de* $\mathbb{S}^1$ *tendría medida nula.*

A partir del push-forward podemos definir un objeto crucial en la formulación de Monge del problema de transporte óptimo: el transporte entre dos medidas.

Definición 1.17 (Transporte entre medidas). *Dados* $\mathcal{X}$ *e* $\mathcal{Y}$ *espacios métricos compactos,* μ *y* ν *medidas en* $\mathcal{X}$ *y* $\mathcal{Y}$ *respectivamente, un transporte* T *entre* μ *y* ν *es una función medible* $T : \mathcal{X} \to \mathcal{Y}$ *que cumple*

$$T_\sharp\mu = \nu.$$

Ejemplo 1.18. *Sea* T *un difeomorfismo* $\mathcal{C}^1$ *entre dos dominios compactos* $\mathcal{X}$ *e* $\mathcal{Y}$ *en* $\mathbb{R}^d$*. Supongamos que* $\mu \in \mathcal{P}(\mathcal{X})$ *y* $\nu \in \mathcal{P}(\mathcal{Y})$ *tienen densidades continuas*

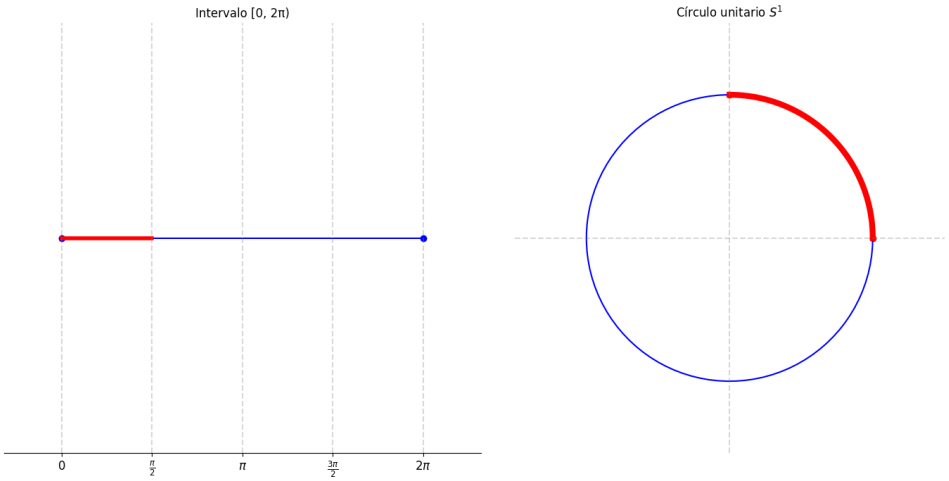


Figura 1.2: A la izquierda el intervalo $[0, 2\pi)$. A la derecha el círculo unitario S^1. El arco de circunferencia rojo va de 0 hasta $\frac{\pi}{2}$ radianes y es la imagen del intervalo rojo mediante la función $T(x) = (cos(x), sen(x))$.

respecto a la medida de Lesbegue ρ y σ respectivamente. Entonces

$$\langle \varphi, \nu \rangle = \int_{\mathcal{Y}} \varphi(y)\sigma(y)dy = \int_{\mathcal{X}} \varphi(T(x))\sigma(T(x)) \det(J_T(x))dx,$$

por lo tanto T es un transporte entre μ y ν si y sólo si para toda $\varphi \in \mathcal{C}^0(\mathcal{Y})$ se tiene que

$$\langle \varphi, \nu \rangle = \int_{\mathcal{X}} \varphi(T(x))\sigma(T(x)) \det(J_T(x))dx = \int_{X} \varphi(T(x))\rho(x)dx = \langle \varphi \circ T, \mu \rangle,$$

donde $J_T(x)$ es la matriz Jacobiana de T.

El ejemplo anterior nos dice que si μ y ν tienen densidades continuas, entonces un transporte T entre μ y ν es equivalente a un cambio de variable entre $\mathcal{X}$ e $\mathcal{Y}$.

Definición 1.19 (Soporte de una medida no negativa). *Dada una medida no negativa $\mu \in \mathcal{M}^+(\mathcal{X})$, decimos que x pertenece al soporte de μ si y sólo si para todo $r > 0$ se tiene que $\mu(B(x, r)) > 0$, donde $B(x, r) = \{z \in \mathcal{X} \mid d(z, x) < r\}$ es la bola en $\mathcal{X}$ de centro x y radio r. El soporte de μ lo denotamos* $\text{sop}(\mu)$.

Observación 1.20. *Se tiene que por el teorema de Banach-Alaoglu, el conjunto de medidas de probabilidad $\mathcal{P}(\mathcal{X})$ es débilmente compacto y por lo tanto toda sucesión tiene al menos una subsucesión convergente. Esto nos ayudara*

más adelante a probar la existencia de soluciones al transporte óptimo en la formulación de Kantorovich.

Un estudio detallado y una demostración del teorema de Banach-Alaoglu puede encontrarse en el libro "Functional Analysis" de Rudin (1991).

Así como el transporte entre medidas es un ingrediente fundamental de la formulación del transporte óptimo según Monge, las medidas productos y sus marginales son la contraparte en la formulación de Kantorovich. A continuación presentamos una serie de definiciones que serán de utilidad en la formulación de Kantorovich.

Definición 1.21 (Proyecciones). *Partiendo del espacio producto $\mathcal{X} \times \mathcal{Y}$ funciones proyección a cada componente son*

$$Pr_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \quad dado\ por \quad Pr_{\mathcal{X}}(x, y) = x$$

y

$$Pr_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} \quad dado\ por \quad Pr_{\mathcal{Y}}(x, y) = y.$$

A partir del push-forward y las proyecciones podemos definir las medidas marginales.

Definición 1.22 (Medidas marginales). *Dada una medida $\gamma \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ llamaremos medidas marginales a las medidas resultantes de aplicar el push-forward a las proyecciones, es decir:*

$$Pr_{\mathcal{X}\sharp}\gamma \in \mathcal{P}(\mathcal{X}) \quad y \quad Pr_{\mathcal{Y}\sharp}\gamma \in \mathcal{P}(\mathcal{Y}).$$

En varias ocasiones a lo largo de este trabajo tendremos que construir una medida en el espacio producto $\mathcal{X} \times \mathcal{Y}$, una forma de hacerlo es con la medida producto, definida a continuación.

Definición 1.23 (Medida producto). *Dada dos medidas $\mu \in \mathcal{M}^+(\mathcal{X})$ y $\nu \in \mathcal{M}^+(\mathcal{Y})$, la medida producto asociada a μ y ν es la medida $\mu \otimes \nu \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y})$ definida como*

$$\mu \otimes \nu(A, B) = \mu(A)\nu(B) \quad para\ todos\ los\ A \subset \mathcal{X} \quad y \quad B \subset \mathcal{Y} \quad medibles.$$

Por otro lado, en ciertas ocasiones nos interesará descomponer una medida μ definida en un espacio producto $\mathcal{X} \times \mathcal{Y}$, como una integral contra alguna medida marginal:

Definición 1.24 (Descomposición de medidas). *Sea $\mu \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ una medida en el espacio producto $\mathcal{X} \times \mathcal{Y}$. Si μ_1 es su marginal en $\mathcal{X}$ entonces existe una única medida μ_x soportada en $\mathcal{Y}$ tal que para todo conjunto medible $A \subset \mathcal{X} \times \mathcal{Y}$ se tiene que*

$$\mu(A) = \int_{\mathcal{X}} \mu_x(A_x) d\mu_1(x) \quad con \quad A_x = \{y \in \mathcal{Y} : (x, y) \in A\}.$$

Escribiremos lo anterior de forma más compacta como

$$\mu = \int_{\mathcal{X}} \delta_x \otimes \mu_x d\mu_1(x).$$

De forma análoga podemos descomponer μ en una integral contra la marginal $\mu_2 \in \mathcal{M}^+(\mathcal{Y})$, obteniendo así que

$$\mu = \int_{\mathcal{Y}} \delta_y \otimes \mu_y d\mu_2(y).$$

Utilizando la descomposición de medidas podemos probar el siguiente lema establecido en el capítulo 1 del libro de Cédric Villani (2009). Nos será de utilidad para probar que la distancia de Wasserstein es, efectivamente, una distancia.

Lema 1.25 (Gluing lemma). *Sean μ_1, μ_2 y μ_3 tres medidas soportadas en tres espacios métricos completos y separables $\mathcal{X}_1, \mathcal{X}_2$ y $\mathcal{X}_3$. Si $\gamma_{1,2} \in \Pi(\mu_1, \mu_2)$ y $\gamma_{2,3} \in \Pi(\mu_2, \mu_3)$ entonces existe una medida γ soportada en $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ tal que tiene marginales $\gamma_{1,2} \in \mathcal{X}_1 \times \mathcal{X}_2$ y $\gamma_{2,3} \in \mathcal{X}_2 \times \mathcal{X}_3$.*

Demostración. La idea principal es factorizar las medidas $\gamma_{1,2}$ y $\gamma_{2,3}$ por su marginal común, es decir

$$\gamma_{1,2} = \int_{\mathcal{X}_2} \gamma_{1,2:\mu_2} \otimes \delta_x d\mu_2(x) \quad \text{y} \quad \gamma_{2,3} = \int_{\mathcal{X}_2} \delta_x \otimes \gamma_{2,3:\mu_2} d\mu_2(x).$$

Luego, construimos una medida π en $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$ 'pegando' las dos anteriores

$$\gamma = \int_{\mathcal{X}_2} \gamma_{1,2:\mu_2} \otimes \delta_x \otimes \gamma_{2,3:\mu_2} d\mu_2(x).$$

Observamos que

$$\begin{aligned}\int_{\mathcal{X}_1} \gamma d\mu_1(x) &= \int_{\mathcal{X}_1} \Big(\int_{\mathcal{X}_2} \gamma_{1,2:\mu_2} \otimes \delta_x \otimes \gamma_{2,3:\mu_2} d\mu_2(x) \Big) d\mu_1(x) \\ &= \int_{\mathcal{X}_2} \delta_x \otimes \gamma_{2,3:\mu_2} d\mu_2(x) \\ &= \gamma_{2,3},\end{aligned}$$

donde en la segunda igualdad utilizamos el teorema de Fubini para intercambiar el orden de integración. De forma análoga $\int_{\mathcal{X}_3} \gamma d\mu_3(x) = \gamma_{1,2}$. □

A continuación introducimos un concepto clave para cuando trabajemos con el problema de transporte óptimo regularizado: la divergencia de Kullback-Leibler, la cual es una manera de cuantificar las diferencias entre dos medidas.

Definición 1.26 (Divergencia Kullback-Leibler). *Sea $\mathcal{X}$ un espacio métrico compacto y $\mu, \nu \in \mathcal{P}(\mathcal{X})$ dos medidas de probabilidad. La divergencia de Kullback-Leibler entre μ y ν es*

$$D_{KL}(\mu||\nu) = \int_{\mathcal{X}} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu.$$

Por convención, si μ no tiene una densidad $\frac{d\mu}{d\nu}$ con respecto a ν diremos que $D_{KL}(\mu||\nu) = +\infty$. En particular, este es el caso si las medidas tienen soporte disjunto.

Observación 1.27. *Cuando tanto μ como ν sean medidas absolutamente continuas respecto a la medida de Lebesgue tenemos que $D_{KL}(\mu||\nu) = \mathbb{E}_\mu\Big(log\Big(\frac{d\mu}{d\nu}\Big)\Big)$.*

Cuando las medidas μ y ν son absolutamente continuas, es decir, $d\mu = f(x)dx$ y $d\nu = g(x)dx$, tenemos que $D_{KL}(\mu||\nu) = \int_{\mathcal{X}} \log\Big(\frac{f(x)}{g(x)}\Big) f(x)dx$, lo cual coincide con la definición usual de la divergencia de Kullback-Leibler.

En el caso discreto, si X e Y son dos variables aleatorias que toman valores en $\mathcal{X} = \{x_1, \ldots, x_n\}$ con densidades p y q entonces la divergencia de Kullback-Leibler es:

$$\sum_{i=1}^{n} p_i \log\Big(\frac{p_i}{q_i}\Big).$$

Una propiedad útil de la divergencia de Kullback-Leibler es que es no negativa, como probaremos a continuación.

Proposición 1.28. *Dadas μ y ν dos medidas de probabilidad en un espacio métrico compacto $\mathcal{X}$ se tiene que*

$$D_{KL}(\mu||\nu) \geq 0,$$

y además

$$D_{KL}(\mu||\nu) = 0 \quad \textit{si y sólo si} \quad \mu = \nu \quad \mu - ctp.$$

Demostración. Para comenzar observamos que $D_{KL}(\mu||\nu) \geq 0$ es equivalente a $-D_{KL}(\mu||\nu) \leq 0$, continuaremos a partir de esta última expresión. Tenemos

$$\begin{aligned}
-D_{KL}(\mu||\nu) &= -\int_{\mathcal{X}} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu \\
&= \int_{\mathcal{X}} \log\Big(\frac{d\nu}{d\mu}\Big) d\mu \\
&= \mathbb{E}_{\mu}\Big(\log\Big(\frac{d\nu}{d\mu}\Big)\Big) \\
&\leq \log\Big(\mathbb{E}_{\mu}\Big(\frac{d\nu}{d\mu}\Big)\Big) \\
&= \log\Big(\int_{\mathcal{X}} \frac{d\nu}{d\mu} d\mu\Big) \\
&= \log(1) \\
&= 0,
\end{aligned}$$

donde en la segunda igualdad utilizamos que $-\log\left(\frac{d\mu}{d\nu}\right) = \log\left(\frac{d\nu}{d\mu}\right)$ y la desigualdad se debe a la desigualdad de Jensen para funciones cóncavas.

Para la segunda afirmación, si llamamos A al conjunto donde $d\mu \neq d\nu$, entonces podemos escribir

$$\begin{aligned}
D_{KL}(\mu||\nu) &= \int_{\mathcal{X}} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu \\
&= \int_{\mathcal{X}\backslash A} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu + \int_{A} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu.
\end{aligned}$$

La primera de estas integrales es 0 ya que en $\mathcal{X}\backslash A$ tenemos que $d\mu = d\nu$ y por lo tanto el logaritmo se anula. Por otro lado, la segunda integral es nula ya que $\mu(A) = 0$ y por convención toda integral sobre un conjunto de medida

nula es cero. □

Es importante notar que a pesar de que la divergencia de Kullback-Leibler tiene estas buenas propiedades, no es una función simétrica, por lo que no es una distancia. Veamos un ejemplo sencillo para ganar intuición sobre el significado de esta divergencia.

Ejemplo 1.29. *Sean $X \sim \eta = \mathcal{N}(\mu_1, \sigma_1)$ y $Y \sim \xi = \mathcal{N}(\mu_2, \sigma_2)$ dos variables aleatorios normales univariadas.*

$$\begin{aligned}\log\Big(\frac{d\eta}{d\xi}\Big) &= \log d\eta - \log d\xi \\ &= -\log(\sqrt{2\pi\sigma_1^2}) - \frac{(x-\mu_1)^2}{2\sigma_1^2} - \Big[-\log(\sqrt{2\pi\sigma_2^2}) - \frac{(x-\mu_2)^2}{2\sigma_2^2}\Big] \\ &= \log\Big(\frac{\sigma_2}{\sigma_1}\Big) + \frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}.\end{aligned}$$

Luego

$$\begin{aligned}\mathbb{E}_\eta\Big(\log\Big(\frac{d\eta}{d\xi}\Big)\Big) &= \mathbb{E}_\eta\Big(\log\Big(\frac{\sigma_2}{\sigma_1}\Big)\Big) + \mathbb{E}_\eta\Big(\frac{(x-\mu_2)^2}{2\sigma_2^2}\Big) - \mathbb{E}_\eta\Big(\frac{(x-\mu_1)^2}{2\sigma_1^2}\Big) \\ &= \log\Big(\frac{\sigma_2}{\sigma_1}\Big) + \frac{1}{2\sigma_2^2}\mathbb{E}_\eta\Big((x-\mu_2)^2\Big) - \frac{1}{2\sigma_1^2}\underbrace{\mathbb{E}_\eta\Big((x-\mu_1)^2\Big)}_{\sigma_1^2} \\ &= \log\Big(\frac{\sigma_2}{\sigma_1}\Big) + \frac{1}{2\sigma_2^2}\mathbb{E}_\eta\Big((x-\mu_2)^2\Big) - \frac{1}{2}.\end{aligned}$$

Por otro lado,

$$(x-\mu_2)^2 = (x-\mu_1+\mu_1-\mu_2)^2 = (x-\mu_1)^2 + 2(x-\mu_1)(\mu_2-\mu_1) + (\mu_1-\mu_2)^2,$$

de donde

$$\begin{aligned}\mathbb{E}_\eta\Big((x-\mu_2)^2\Big) &= \mathbb{E}_\eta\Big((x-\mu_1)^2\Big) + \mathbb{E}_\eta\Big(2(x-\mu_1)(\mu_2-\mu_1)\Big) + \mathbb{E}_\eta\Big((\mu_1-\mu_2)^2\Big) \\ &= \sigma_1^2 + 2(\mu_2-\mu_1)\underbrace{\mathbb{E}_\eta\Big((x-\mu_1)\Big)}_{0} + (\mu_1-\mu_2)^2 \\ &= \sigma_1^2 + (\mu_1-\mu_2)^2.\end{aligned}$$

Juntando todo tenemos

$$D_{KL}(\eta||\xi) = log\Big(\frac{\sigma_2}{\sigma_1}\Big) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

Ejemplo 1.30. *En el caso particular donde* $X \sim \eta = \mathcal{N}(\mu_1, \sigma)$ *e* $Y \sim \xi = \mathcal{N}(\mu_2, \sigma)$ *tienen misma varianza tenemos*

$$D_{KL}(\eta||\xi) = \frac{(\mu_1 - \mu_2)^2}{\sigma^2}.$$

Es interesante notar que dejando fijas las medias μ_1 *y* μ_2 *podemos hacer tender a* 0 *o* ∞ *la divergencia* $D_{KL}(\eta||\xi)$*. El caso donde* $\sigma \to 0$*, el cuál hace que la divergencia* KL *tienda a* ∞*, coincide con la definición de la misma, ya que en este caso* η *y* ξ *tienen soporte disjuntos.*

La divergencia de Kullback-Leibler D_{KL} posee una relación estrecha y significativa con la función de log verosimilitud. Esta relación se manifiesta en el contexto de la estimación de parámetros de modelos estadísticos y nos permite entender cómo la minimización de una implica la maximización de la otra.

La función de verosimilitud se utiliza para estimar los parámetros de un modelo de probabilidad, representando la probabilidad de observar los datos dados ciertos parámetros. Al tomar el logaritmo de esta función, obtenemos la verosimilitud logarítmica, que es más manejable matemáticamente. Explícitamente, si $\{x_1, \ldots, x_n\}$ es un conjunto de datos el cual queremos ajustar con un modelo paramétrico con densidad $q_\theta(x)$ que depende de θ, la función de verosimilitud es es

$$L(\theta|x_1, \ldots, x_n) = q_\theta(x_1, \ldots, x_n) = \Pi_{i=1}^n q_\theta(x_i),$$

que al tomar logaritmo queda

$$\ell(\theta|x_1, \ldots, x_n) = \log L(\theta|x_1, \ldots, x_n) = \sum_{i=1}^{n} \log q_\theta(x_i),$$

y se busca encontrar θ que maximiza dicha expresión.

Por otro lado, si suponemos que la densidad verdadera es $p(x)$ entonces podemos utilizar la divergencia de Kullback-Leibler para buscar el parámetro

que mejor se ajuste, es decir,

$$\theta^* = \arg\min_{\theta} D_{KL}(p(x)||q_\theta(x)).$$

Recordando la definición de D_{KL} tenemos

$$\begin{aligned} D_{KL}(p(x)||q_\theta(x)) &= \int p(x) \log\Big(\frac{p(x)}{q_\theta(x)}\Big) dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q_\theta(x) dx, \end{aligned}$$

como el primer sumando no depende de θ al momento de minimizar podemos quedarnos únicamente con la segunda integral de la expresión anterior, la cual además, se puede aproximar de la siguiente forma:

$$-\int p(x) \log q_\theta(x) dx \approx -\frac{1}{n} \sum_{i=1}^{n} \log q_\theta(x_i),$$

por lo que minimizar la divergencia de Kullback-Leibler entre $p(x)$ y $q_\theta(x)$ es equivalente a maximizar la verosimilitud $\sum_{i=1}^{n} \log q_\theta(x_i)$.

Cuando trabajemos en el caso discreto, remplazaremos la divergencia de Kullback-Leibler por la entropía, la cual se define como

Definición 1.31 (Entropía). *Dada una variable aleatoria discreta X que toma valores en $\mathcal{X} = \{x_1, \ldots, x_n\}$ con densidad p (es decir, $P(X = x_i) = p_i$) se define la entropía de X como*

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i.$$

La entropía es una medida de la incertidumbre asociada a una variable aleatoria. A partir de la entropía podemos definir lo que usualmente se conoce como entropía cruzada

Definición 1.32 (Entropía cruzada). *Si X e Y son variables aleatorias que toman valores en $\mathcal{X} = \{x_1, \ldots, x_n\}$ con densidades p y q respectivamente, se define la entropía cruzada como*

$$H(X, Y) = -\sum_{i=1}^{n} p_i \log q_i.$$

Existe una relación clara entre la divergencia de Kullback-Leibler y la entropía: sean X e Y dos variables aleatorias discretas que toman valores en $\mathcal{X} = \{x_1, \ldots, x_n\}$ con densidades p y q respectivamente, entonces

$$\begin{aligned} D_{KL}(X||Y) &= \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \\ &= \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i \\ &= -H(X) + H(X,Y). \end{aligned}$$

1.2. Operaciones matriciales

A continuación, introducimos las matrices bistocásticas, las cuales serán de utilidad para uno de los algoritmos que presentaremos en el capítulo 3.2. Estas matrices, que tienen la propiedad de que sus filas y columnas suman uno, son fundamentales para el problema de asignación lineal.

Definición 1.33 (Matrices Bistocásticas). *Decimos que una matriz cuadrada M de tamaño $N \times N$ es bistocásticas si sus entradas son no negativas y además la suma de cualquier fila o columna es* 1*, es decir,*

$$\sum_{i=1}^{N} M_{ij} = 1, \quad \sum_{j=1}^{N} M_{ij} = 1, \quad \forall i, j \in \{1, \ldots, N\}.$$

Denotaremos el conjunto de matrices bistocásticas de tamaño $N \times N$ como $\mathcal{B}_N$.

Otro ingrediente necesario para el problema de asignación lineal son las matrices de permutación. Estas son matrices cuadradas que representan permutaciones de elementos en un conjunto, y por lo tanto todas las entradas de cualquier fila o columna son nulas, excepto una que vale 1. Estas matrices son ampliamente usadas en problemas de optimización combinatoria para reordenar elementos en un conjunto.

Definición 1.34 (Matrices de permutación). *Dada una permutación σ del conjunto $\{1, \ldots, N\}$ la matriz de permutación asociada a σ es una matriz M_σ*

de tamaño $N \times N$ donde las entradas son

$$((M_\sigma))_{ij} = \begin{cases} 1 \text{ } si \text{ } \sigma(i) = j, \\ 0 \text{ } si \text{ } no. \end{cases}$$

Denotaremos el conjunto de matrices de permutación de $\{1, \ldots, N\}$ por $\mathcal{G}_N$.

Es claro que $\mathcal{G}_N \subset \mathcal{B}_N$. El siguiente teorema afirma que los vértices del poliedro $\mathcal{B}_N$ son las matrices de permutaciones, lo cual por el teorema de Krein-Milman implica que cualquier matriz bistocásticas se puede obtener como una combinación convexa finita de matrices de permutación.

Teorema 1.35 (Teorema de Birkhoff). *Los vértices del poliedro $\mathcal{B}_N$ son las matrices de permutación. En particular, $\mathcal{B}_N = conv\{M[\sigma] \mid \sigma \in \mathcal{G}_N\}$.*

El teorema de Birkhoff es un resultado clásico en el área de optimización matemática, una demostración de dicho teorema puede encontrarse en el libro "Matrix Analysis" de Horn y Johnson (2013).

Al trabajar con los problemas discretos, sera usual utilizar producto entre matrices de igual dimensión, para simplificar la notación introducimos el producto interno de Frobenius, que no es otra cosa que el producto punto a punto entre matrices.

Definición 1.36 (Producto interno de Frobenius). *Dadas dos matrices $A = (a_{ij})_{ij}$ y $B = (b_{ij})_{ij} \in \mathcal{M}_{n\times m}$, su producto interno de Frobenius es*

$$\langle A, B\rangle_F = \sum_{i,j} A_{i,j} B_{i,j}.$$

Por otro lado, cuando trabajemos con el algoritmo de Sinkhorn-Knopp nos será de utilidad la siguiente notación: si $\mathbf{a}$ y $\mathbf{b}$ dos vectores en $\mathbb{R}^n$ denotaremos $\mathbf{a} \odot \mathbf{b}$ al producto punto a punto de los vectores $\mathbf{a}$ y $\mathbf{b}$, es decir

$$\mathbf{a} \odot \mathbf{b} = (a_i b_i)_i \in \mathbb{R}^n.$$

Si $b_i \neq 0$ para todo $i = 1, \ldots, n$, notaremos $\mathbf{a} \oslash \mathbf{b}$ al cociente punto a punto, es decir

$$\mathbf{a} \oslash \mathbf{b} = \left(\frac{a_i}{b_i}\right)_i \in \mathbb{R}^n.$$

Por último, introducimos la descomposición en valores singulares (SVD), la cual es una técnica que nos servirá para estimar los ángulos entre dos conjuntos de datos, lo que usaremos en una de las simulaciones.

Definición 1.37. *Decimos que una matriz $A \in \mathbb{R}^{n\times n}$ es ortogonal si*

$$A^T A = Id.$$

Teorema 1.38 (Descomposición en valores singulares). *Toda matriz $A \in \mathbb{R}^{m\times n}$ puede descomponerse de la forma*

$$A = U \cdot \Sigma \cdot V^T,$$

donde U y V son matrices ortogonales y Σ es una matriz diagonal.

Las entradas no nulas de Σ se denominan valores singulares.

En una de las simulaciones supondremos que tenemos dos conjuntos de puntos relacionados mediante un rotación. Presentamos a continuación un procedimiento para estimar el ángulo de dicha rotación utilizando la descomposición en valores singulares, el mismo fue propuesto en el artículo "Least-Squares Fitting of Two 3-D Point Sets" de Arun et al. (1987).

Sean $X = \{x_1, \ldots, x_n\}$ e $Y = \{y_1, \ldots, y_n\}$ dos conjuntos de puntos en el plano relacionados mediante una rotación. Abusando de la notación podemos escribir $Y = RX$. El objetivo es encontrar una estimación $\hat{R}$ de la matriz R a partir de X e Y. El procedimiento consiste en:

1. Calcular $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ e $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$.
2. Calcular $\hat{x_j} = x_j - \overline{x}$ e $\hat{y_j} = y_j - \overline{y}$ para todo $j = 1, \ldots, n$.
3. Calcular la matriz H como $\sum_{j=1}^{n} \hat{x_j}\hat{y_j}^T$.
4. Encontrar la descomposición en valores singulares: $H = U\Sigma V^T$.
5. Definimos $Z = VU^T$.
6. Si el determinante de Z es $+1$ entonces $\hat{R} = Z$. Si el determinante de Z es -1 entonces se cambia el signo a la última fila de V y se recalcula $Z = VU^T = \hat{R}$.

Si R es una matriz de rotación en el plano entonces la primera columna es un vector de la forma $z = (cos\theta, sen\theta)$, luego el ángulo de rotación es el argumento de z visto como número complejo.

Capítulo 2

Transporte óptimo

El transporte óptimo es una teoría matemática que ha captado la atención de investigadores de diversas disciplinas debido a su capacidad para resolver problemas complejos de asignación y distribución de recursos. Sus comienzos datan del siglo XVIII y hasta el día de hoy sigue siendo un tema activo de investigación. La figura 2.1 muestra algunos de los matemáticos que realizaron contribuciones notables a la teoría a lo largo del tiempo, comenzando con Gaspard Monge (Monge, 1781), siguiendo por Leonid Kantorovich (Kantorovich, 1942) quien además gano el premio Nobel en Economía por dicho trabajo en el año 1975. En el año 1991 Brenier mostró la existencia de soluciones al problema de transporte óptimo en algunos casos particulares mientras que en el siglo XXI McCann, Villani y Figalli incursionaron en la geometría subyacente. Una excelente referencia sobre una descripción completa del tema es el libro de Villani (2009).

Este capítulo tiene como objetivo proporcionar una base sólida en los fundamentos matemáticos del transporte óptimo, así como explorar sus aplicaciones en el campo del aprendizaje automático. A lo largo de las siguientes secciones,

Figura 2.1: Linea de tiempo sobre los avances más notables en Transporte Óptimo. Imagen tomada de remi.flamary.com.

se discutirán las formulaciones clásicas del problema de transporte óptimo y se examinarán diversas metodologías para su solución.

En la primera sección presentaremos dos enfoques clásicos del problema de transporte óptimo: la formulación original de Monge (1781) y la relajación propuesta por Kantorovich (1942). Introduciremos la distancia de Wasserstein, una métrica clave en este campo. En la segunda sección estudiaremos el problema dual de la formulación de Kantorovich del transporte óptimo ya que será fundamental para estudiar luego el problema del transporte óptimo regularizado. Finalmente, en la tercera sección discutiremos el problema de transporte óptimo regularizado, el cual utiliza la divergencia de Kullback-Leibler. En esta sección además compararemos la distancia de Wasserstein con la divergencia de Kullback-Leibler.

Este capítulo pretende no solo proporcionar una comprensión teórica profunda del transporte óptimo, tanto en la formulación de Monge como en la de Kantorovich, sino también introducir la distancia de Wasserstein y el problema del transporte óptimo regularizado, que constituye el avance más reciente de la teoría.

2.1. Fundamentos teóricos

En esta sección, se presentarán los conceptos teóricos fundamentales que forman las bases del transporte óptimo. Primero, introduciremos la formulación de Monge, la primera aproximación a la teoría del transporte óptimo. Esta formulación se centra en la asignación de recursos de un lugar a otro de manera a minimizar el costo del transporte. Exploraremos los conceptos claves y las condiciones necesarias para la existencia de soluciones óptimas dentro de este marco.

A continuación, discutiremos la formulación de Kantorovich, introducida casi 200 años después, que extiende y generaliza el enfoque de Monge. Esta formulación introduce una perspectiva más flexible y robusta, permitiendo la consideración de planes de transporte y una mejor adaptabilidad a una variedad de aplicaciones prácticas. Se explicarán los fundamentos matemáticos y las ventajas que ofrece esta perspectiva.

Por último, abordaremos el concepto de la distancia de Wasserstein, una métrica clave en el análisis de distribuciones de probabilidad y en la teoría

Figura 2.2: Cuadro de Gaspard Monge (1746-1818), el mismo se encuentra en el Museo de la Historia de Francia (Château de Versailles). Imagen tomada de es.wikipedia.org.

del transporte óptimo. Se explicará cómo esta distancia cuantifica las diferencias entre distribuciones y se describirán sus propiedades fundamentales y aplicaciones en diversos contextos.

2.1.1. Formulación de Monge

Gaspard Monge (1746-1818), matemático francés del siglo XVIII, es conocido como el fundador de la teoría del transporte óptimo. Monge estaba interesado en resolver problemas prácticos de ingeniería y economía, específicamente en la manera más eficiente de mover tierra o materiales de un lugar a otro, minimizando el costo total del transporte. La figura 2.2 es un retrato de Gaspard Monge.

En su obra "Mémoire sur la théorie des déblais et des remblais"(Memoria sobre la teoría de los desechos y los rellenos) (Monge, 1781), Monge se enfrentó al problema de cómo trasladar una cantidad de arena de una estructura inicial a una estructura final de la manera más eficiente posible. Se dio cuenta de que este tipo de problemas podía ser modelado matemáticamente, lo que permitía encontrar soluciones óptimas mediante el uso de técnicas geométricas y analíticas. Su trabajo sentó las bases para la teoría moderna del transporte óptimo, que ha encontrado aplicaciones en diversas áreas como la economía, la logística, la teoría de juegos, y más recientemente, en el aprendizaje automático y la ciencia de datos.

Podemos modelar la pila de tierra de origen y de destino con medidas de probabilidad P y Q mientras que el costo asociado proviene de la distancia entre las posiciones de origen y de destino d. La figura 2.3 bosqueja la situación. La forma de 'mover' la tierra estará dada por un mapa T entre estas medidas que para cada x de origen nos dirá la posición a ocupar $T(x)$ al final.

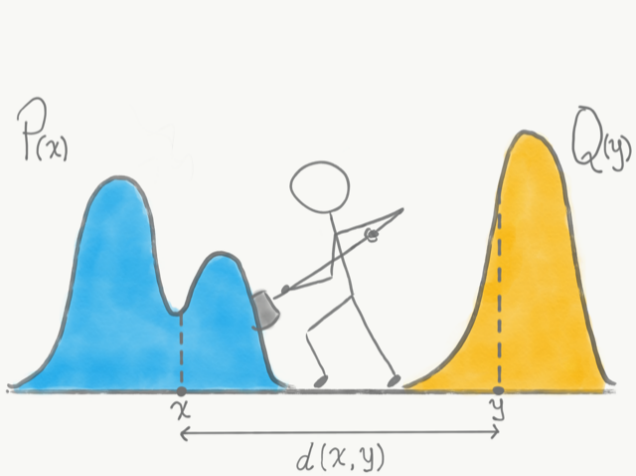


Figura 2.3: Bosquejo del problema de Monge. Imagen tomada de www.microsoft.com.

Con estas definiciones, el problema de Monge consiste en encontrar el mapa T que minimiza el costo total del transporte

$$\int d(x, T(x))dP(x),$$

con la restricción de que toda la masa proveniente de una pila de arena sea transportada vía T a la estructura final.

En un contexto más formal, tendremos dos espacios métricos compactos $\mathcal{X}$ e $\mathcal{Y}$, escribiremos las medidas de probabilidad como $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$ y para representar el costo de transportar el material utilizamos una función continua $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ la cuál llamaremos, como era de esperar, función de costo. A partir de esto el costo total de transportar la medida μ a la medida ν se define como

$$\int_{\mathcal{X}} c(x, T(x))d\mu(x).$$

Como lo que queremos es minimizar el costo total, el problema de Monge es

Definición 2.1 (Problema de Monge). *Dados $\mathcal{X}$ e $\mathcal{Y}$ dos espacios métricos compactos, dos medidas de probabilidad $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$ y una función de costo $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ continua, el problema de Monge (MP) es el problema de optimización*

$$T^* \quad = \quad \underset{T_\sharp \mu = \nu}{\arg\min} \Big\{ \int_{\mathcal{X}} c(x, T(x)) d\mu \Big\}. \qquad (MP)$$

Ejemplo 2.2. *Inicialmente Monge propuso el costo $c(x, T(x)) = |x - T(x)|$, es decir, el costo es simplemente la distancia entre el origen y el destino. Si $\mu = \frac{1}{3}\delta_0 + \frac{1}{3}\delta_1 + \frac{1}{3}\delta_{10}$ y $\nu = \frac{2}{3}\delta_3 + \frac{1}{3}\delta_7$ son dos medidas de probabilidad discretas en $\mathbb{R}$ entonces es evidente que si el costo de transporte es la distancia entre el punto de partida y el de llegada, el transporte que minimiza el costo es el que manda las masas de las posiciones 0 y 1 a la posición 3 y la masa de la posición 10 a la posición 7. Formalmente podemos escribir*

$$T(0) = 3, \;\; T(1) = 3 \;\; y \;\; T(10) = 7.$$

La figura 2.4 muestra la situación descrita y el transporte óptimo.

Figura 2.4: En rojo las masas de salida y en azul las de llegada del ejemplo 2.2. El tamaño del círculo representa la masa 'depositada' en el mismo.

El costo asociado a este transporte es

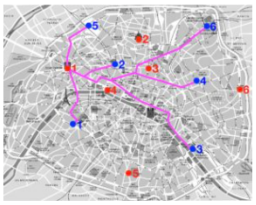
$$\sum_{k=1}^{3} |x_k - T(x_k)| p_x(k) = |0 - 3|\frac{1}{3} + |1 - 3|\frac{1}{3} + |10 - 7|\frac{1}{3} = \frac{8}{3},$$

donde $p_x(k) = \frac{1}{3}$ para $k = 1, 2, 3$.

Presentamos a continuación un ejemplo más ilustrativo sobre lo que busca resolver el problema de Monge, el mismo fue extraído del blog de Gabriel Peyré.

Ejemplo 2.3. *Supongamos para simplificar que en París hay 6 panaderías $\{x_1, \ldots, x_6\}$ y 6 cafeterías $\{y_1, \ldots, y_6\}$ y que cada panadería fabrica la misma*

c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15



c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

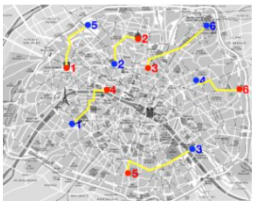


Figura 2.5: Ejemplo de un problema de transporte optimal entre panaderías y cafeterías en Paris. En la matriz de costos (tiempos) está el tiempo que se requiere para ir de la panadería j a la cafetería i. Si cada panadería es proveedora de una sola cafetería, los casilleros amarillos se corresponden a una posible solución candidata a minimizar el tiempo en que se entrega la mercadería. Este ejemplo es tomando del siguiente blog y fue utilizado por Gabriel Peyré.

cantidad de croissants y que cada cafetería solicita también la misma cantidad de croissants. Suponemos además que cada cafetería y_i puede recibir la mercadería de una única panadería x_j. El costo que se quiere minimizar es el tiempo total de viaje entre cada panadería a cada cafetería. La figura 2.5 muestra a la izquierda la matriz de costo (tiempos de traslado) entre panaderías x_j y cafeterías y_i. En el primer renglón (en rosado), tenemos que el tiempo de transportar los croissants de la panadería 1 *a la cafetería* 1 *es de 12 minutos mientras que el tiempo de transportarlos a la cafetería* 2 *es 10 minutos, etc. Sobre el mapa, en rosado, se pueden ver estos trayectos. En la segunda matriz, en amarillo, tenemos un transporte que, con nuestras suposiciones, es una permutación $\sigma = \{1, \ldots, 6\} =$ panaderías $\to \{1, \ldots, 6\} =$ cafeterías con costo $10+7+15+10+14+9 = 65$ y lo que nos preguntamos es si esta permutación es candidata en hacer que el costo global sea mínimo.*

Lo mostrado en el ejemplo anterior no es un caso aislado, de hecho, siempre que $\mathcal{X} = \{x_1, \ldots, x_N\}$ e $\mathcal{Y} = \{y_1, \ldots, y_N\}$ sean conjuntos con la misma cantidad de puntos y $\mu = \sum_{i=1}^{N} \frac{1}{N}\delta_{x_i}, \nu = \sum_{j=1}^{N} \frac{1}{N}\delta_{y_j}$ sean medidas uniformes sobre estos conjuntos, tenemos que el problema de Monge consiste en encontrar la permutación σ que minimiza el costo global, es decir,

$$\sigma^* = \arg\min_{\sigma \in S_N} \left\{ \sum_{i=1}^{N} C_{i\sigma(i)} \right\}.$$

siendo C_{ij} el costo de ir de la panadería i a la cafetería j. Este problema es conocido como el problema de asignación lineal, el cual puede ser resuelto por ejemplo con el algoritmo 1 del capítulo 3 (algoritmo Húngaro).

A pesar de que la formulación de Monge es intuitiva tiene un gran inconveniente desde el punto de vista de la optimización: la restricción $T_\sharp\mu = \nu$ hace que el problema de optimización no siempre tenga solución, como muestra el siguiente ejemplo.

Ejemplo 2.4. *Supongamos ahora que* $\mu = \delta_5$ *y* $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{10}$. *Para poder mandar la masa concentrada en la posición* 5 *a las otras dos, necesitamos una función sobreyectiva* $T : \{5\} \to \{0, 10\}$, *pero tal función no existe por ser los conjuntos dominio y codominio finitos y este último tener un cardinal mayor que el primero. Por lo tanto no existe ningún transporte. La figura 2.6 muestra la situación.*

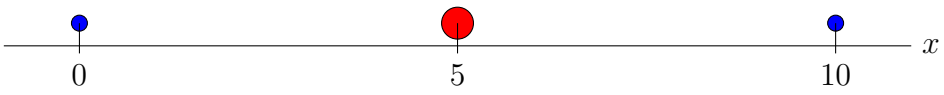


Figura 2.6: En rojo la masa de salida y en azul las de llegada. El tamaño del círculo representa la masa 'depositada' en el mismo.

El ejemplo anterior podría decirse que es patológico, ya que no existe ninguna función T candidata a ser solución. Veamos a continuación un ejemplo extraído del trabajo de Bruno Levy (2014), que muestra que, aunque haya funciones candidatas, no tiene por qué haber un óptimo, y por lo tanto no tiene por que haber solución al problema de Monge.

Ejemplo 2.5. *Sean* $L_2 = \{(-1, y) \mid y \in [0, \frac{1}{2}]\}$, $L_1 = \{(0, y) \mid y \in [0, \frac{1}{2}]\}$ *y* $L_3 = \{(1, y) \mid y \in [0, \frac{1}{2}]\}$. *Consideramos las medidas*

$$\mu = 2\mathbf{1}_{L_1}(x) \quad y \quad \nu = \mathbf{1}_{L_2}(x) + \mathbf{1}_{L_3}(x),$$

donde $\mathbf{1}_{L_j}(x)$ *es la función indicatriz en* L_j. *Utilizamos además como función de costo la distancia euclídea.*

Sea $h > 0$ *tal que existe* $n \in \mathbb{N}$ *que cumple* $nh = \frac{1}{2}$ *(por ejemplo* $h = \frac{1}{4}$ *y* $n = 2$ *o* $h = \frac{1}{12}$ *y* $n = 6$*). Fijado* h, *dividimos el intervalo* L_1 *de la siguiente forma:*

$$L_k = \Big\{(0, y) \mid y \in [(k-1)h, kh)\Big\} \text{ para } k = 1, \dots, n.$$

Luego podemos construir un transporte T *de la siguiente manera:*

$$T(L_k) = \begin{cases} \Big\{(1, y) \mid y \in [(k-1)h, (k+1)h)\Big\} \subset L_3 & \text{si } k \text{ es impar,} \\ \Big\{(-1, y) \mid y \in [(k-2)h, kh)\Big\} \subset L_2 & \text{si } k \text{ es par.} \end{cases}$$

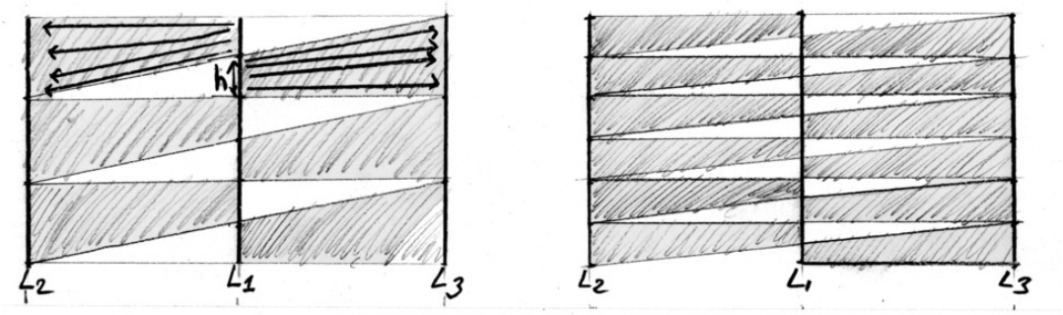


Figura 2.7: Bosquejo de la situación del ejemplo 2.5. Ambas son soluciones admisibles, pero la de la derecha tiene menor costo. El proceso se puede continuar indefinidamente, por lo que no existe un plan de transporte óptimo. Imagen tomada de Levy, 2014.

La figura 2.7 muestra un bosquejo de la situación. Luego, si tomamos $h' = \frac{h}{2}$ *y repetimos el procedimiento obtenemos otro mapa* T' *pero ahora el costo es menor, dado que la distancia vertical que hay que 'mover' cada unidad de masa es menor. Este procedimiento se puede continuar indefinidamente, encontrar así en cada iteración una nueva solución admisible cuyo costo es estrictamente inferior a todos las soluciones admisibles en los pasos anteriores.*

Finalmente, si pensamos el límite de este procedimiento como el límite cuando h *tiende a cero, obtenemos que cada punto de* L_2 *se tiene que separar en dos partes de igual peso (split de masa) transportando la mitad para* L_1 *y la otra mitad para* L_2*, pero esto no es un transporte en el sentido de Monge.*

Aún en los casos donde exista un transporte óptimo, no tiene por que ser único, como muestra el siguiente ejemplo.

Ejemplo 2.6. *La figura 2.8 representa la situación donde* $\mathcal{X} = \{x_1, x_2\} = \{(0,0),(1,1)\}$ *e* $\mathcal{Y} = \{y_1, y_2\} = \{(0,1),(1,0)\}$ *y supongamos que tenemos dos medidas una soportada en* $\mathcal{X}$*,* $\mu = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)}$ *y otra soportada en* $\mathcal{Y}$*,* $\nu = \frac{1}{2}\delta_{(1,0)} + \frac{1}{2}\delta_{(0,1)}$*. Consideramos la distancia euclídea* $||\cdot||_2$ *como función de costo.*

Un posible transporte es $T_1(x_1) = y_1$ *y* $T_1(x_2) = y_2$*, el cual corresponde con moverse en las lineas punteadas. El costo de este transporte es*

$$\sum_{i=1}^{2} \frac{1}{2}||x_i - T_1(x_i)||_2 = 1,$$

ya que $||x_i - T_1(x_i)||_2 = 1$ *para* $i = 1, 2$.

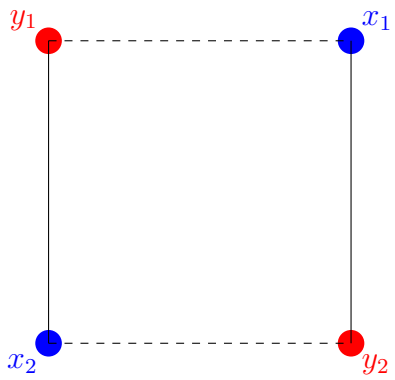


Figura 2.8: En azul las masas de salida y en rojo las de llegada. Existen dos posibles transporte óptimos: moverse por las lineas horizontales o las verticales.

Otro posible transporte es $T_2(x_1) = y_2$ *y* $T_2(x_2) = y_1$*, correspondiente a moverse por las lineas sólidas. El costo de este transporte también es*

$$\sum_{i=1}^{2} \frac{1}{2}||x_i - T_2(x_i)||_2 = 1,$$

ya que $||x_i - T_2(x_i)||_2 = 1$ *para* $i = 1, 2$.

Por otro lado, como el mapa T *tiene que ser una biyección entre* $\mathcal{X} = \{(0,0),(1,1)\}$ *e* $\mathcal{Y} = \{(1,0),(0,1)\}$ *sabemos que solo hay dos posibilidades, es decir, los transportes descritos anteriormente son los únicos posibles.*

Por último, ambos transportes tienen el mismo costo, lo cual implica que ambos son óptimos y por lo tanto no hay unicidad.

A pesar de que la formulación de Monge tiene una interpretación sencilla, tiene algunas dificultades, por ejemplo, que la restricción de que $T_\sharp\mu = \nu$ hace que en general el problema de optimización no tenga solución y que no haya unicidad. En la siguiente sección introduciremos la formulación de Kantorovich, la cual hizo re-aparecer el transporte óptimo luego de casi 200 años.

2.1.2. Formulación de Kantorovich

Aproximadamente dos siglos después de los trabajos de Monge, Leonid Kantorovich (1912-1986), un matemático y economista ruso, retomó el problema del transporte óptimo en la década de 1940 (Kantorovich, 1942). Kantorovich estaba interesado en optimizar recursos en una economía planificada, y su abordaje del transporte óptimo surgió de la necesidad de mejorar la eficiencia

Figura 2.9: Fotografía de Leonid Kantorovich (1912-1986). Imagen tomada de en.scientificrussia.ru.

en la asignación de recursos y en la planificación económica. La figura 2.9 es una fotografía de Kantorovich.

Kantorovich estaba pensando en cómo encontrar la manera más eficiente de asignar recursos limitados en una economía, minimizando los costos de transporte y otros gastos operativos. Su trabajo proporcionó una base matemática sólida para la programación lineal y el análisis de sistemas económicos complejos. Introdujo el concepto de dualidad en la programación lineal y mostró cómo estos métodos podían aplicarse para resolver problemas de transporte y asignación de recursos, lo que revolucionó la forma en que se abordarían estos problemas. Por su trabajo en esta área, Kantorovich fue galardonado con el Premio Nobel de Economía en 1975, compartido con Tjalling Koopmans, por sus contribuciones a la teoría de la asignación óptima de recursos.

Al igual que en la formulación de Monge, consideramos un par de espacios métricos compactos $\mathcal{X}$ e $\mathcal{Y}$ y un par de medidas de probabilidad $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$, pero en lugar de buscar un mapa $T : \mathcal{X} \to \mathcal{Y}$ que sea un transporte de μ a ν, Kantorovich propone buscar una medida γ en el espacio producto $\mathcal{X} \times \mathcal{Y}$ de tal manera que sus marginales en $\mathcal{X}$ e $\mathcal{Y}$ sean μ y ν respectivamente. A continuación introducimos algunos conceptos necesarios para formalizar estas ideas.

Definición 2.7 (Plan de transporte). *Sean $\mathcal{X}$ e $\mathcal{Y}$ dos espacios métricos compactos, $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$. Un plan de transporte γ entre μ y ν es una medida en el espacio producto $\mathcal{X} \times \mathcal{Y}$ cuyas marginales son μ y ν, es decir, $Pr_{\mathcal{X}\sharp}(\gamma) = \mu$ y $Pr_{\mathcal{Y}\sharp}(\gamma) = \nu$. Al espacio de planes de transporte lo detonamos*

por $\Pi(\mu,\nu)$*:*

$$\Pi(\mu,\nu) = \Big\{\gamma \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) \;\; | \;\; Pr_{\mathcal{X}\sharp}(\gamma) = \mu, \;\; Pr_{\mathcal{Y}\sharp}(\gamma) = \nu\Big\}.$$

Observación 2.8. *En la literatura inglesa al plan de transporte* γ *se le suele llamar "coupling".*

Al igual que en la formulación de Monge, tenemos una función de costo $c \in \mathcal{C}^0(\mathcal{X}\times\mathcal{Y})$ no negativa. El costo total queda reformulado de la siguiente manera:

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y).$$

Observación 2.9. *La formulación anterior es la formulación clásica para el costo de Kantorovich, pero tenemos también una interpretación probabilística. Supongamos que* X *e* Y *son dos variables aleatorias en,* $\mathcal{X}$ *e* $\mathcal{Y}$ *respectivamente, con distribución* μ *y* ν*. La condición* $\gamma \in \Pi(\mu,\nu)$ *es equivalente a que la distribución conjunta de* (X,Y) *sea* γ*, es decir,* $(X,Y)\sim\gamma$*. A partir de esto podemos escribir el costo total de Kantorovich como una esperanza*

$$\int_{X\times Y} c(x,y)\,d\gamma(x,y) = \mathop{\mathbb{E}}_{\substack{X\sim\mu\\ Y\sim\nu\\ (X,Y)\sim\gamma}} \Big(c(X,Y)\Big)$$

e interpretar el costo de Kantorovich como el costo medio de transportar μ *a* ν *utilizando el plan de transporte* γ*.*

A pesar de que la formulación de Kantorovich tiene una estrategia distinta a la de Monge, el objetiva final sigue siendo el mismo: encontrar la manera de transportar μ a ν minimizando el costo c.

Definición 2.10 (Problema Kantorovich). *Dados dos espacios métricos compactos* $\mathcal{X}$ *e* $\mathcal{Y}$*, dos medidas de probabilidad* $\mu \in \mathcal{P}(X)$ *y* $\nu \in \mathcal{P}(Y)$ *y una función de costo* $c \in \mathcal{C}^0(\mathcal{X}\times\mathcal{Y})$ *el problema de Kantorovich* (KP) *es el problema de optimización*

$$\mathop{\arg\min}_{\gamma\in\Pi(\mu,\nu)} \Big\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma(x,y)\Big\} \qquad (KP)$$

A partir de la observación 2.9, podemos interpretar el problema de Kantorovich como encontrar un plan de transporte γ^* que minimiza el costo medio

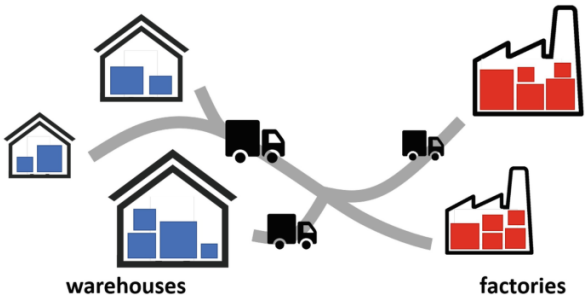


Figura 2.10: Bosquejo de la situación explicada en 2.1.2. Figura tomada de Zhang et al. (2021).

de transportar μ a ν, esto implica que 'localmente' pueden existir otros planes de transporte γ que tengan un costo menor a γ^*, pero globalmente no.

Así como Monge estaba pensando en el problema de transportar materiales de un lugar a otro minimizando la distancia recorrida, Kantorovich estaba pensando también en un problema de asignación de recursos, pero en un contexto más moderno: supongamos que tenemos n depósitos y m fabricas. Cada depósito contiene una materia prima necesaria para el correcto funcionamiento de las fábricas. Indexando cada depósito con i, supongamos que cada uno tiene μ_i unidades de esta materia prima. Por otro lado, indexando las fabricas con j, supongamos que cada una necesita ν_j unidades del material para funcionar correctamente. En la ciudad hay una única compañía de traslados la cual cobra un importe de C_{ij} por trasladar cada unidad de material desde el depósito i a la fabrica j. Suponemos además que el costo crece linealmente, es decir, si queremos transportar μ unidades del depósito i a la fabrica j el costo será μC_{ij}. Si P_γ es un plan de trasporte entonces la compañía de repartos nos cobrará en total $\langle C, P_\gamma \rangle_F$. Nuestro objetivo es encontrar el plan de transporte óptimo que minimiza los costos de traslados.

La utilidad de la formulación de Kantorovich es que, a diferencia del problema de Monge, podemos garantizar la existencia de la solución al problema de transporte óptimo (ver teorema 2.14). Para ejemplificar esto retomamos el ejemplo 2.4.

Ejemplo 2.11. *Sean $\mu = \delta_5$ y $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{10}$. Como vimos en el ejemplo 2.4, no existe un mapa T que resuelva el problema de Monge. Veremos a continuación que existe un plan de transporte que resuelve el problema de Kantorovich en este contexto.*

En el caso discreto, una medida γ en el espacio producto $\mathcal{X} \times \mathcal{Y}$ se puede representar mediante una matriz de tamaño $n \times m$ donde n y m son los cardinales de $\mathcal{X}$ e $\mathcal{Y}$ respectivamente. En nuestro caso, $n = 1$ y $m = 2$ por lo que γ será una matriz 1×2. Explícitamente para este ejemplo queda $\gamma = (\frac{1}{2}, \frac{1}{2})$, lo cual puede interpretarse como asignar la mitad de la masa en la posición 5 a la posición 0 y la otra mitad a la posición 10.

Otro ejemplo sale de tomar el proceso "límite" en el ejemplo 2.5, donde para cada masa del segmento L_2 se transporta la mitad de la misma a L_1 y la otra mitad a L_3. Como mencionamos en ese ejemplo, esto no es un mapa del tipo de Monge, pero si es un plan de transporte, y de hecho, es el plan de transporte óptimo.

A continuación nos dedicaremos a demostrar una serie de propiedades que son fundamentales para asegurar la existencia del plan de transporte óptimo.

Proposición 2.12. *Si μ y ν son dos medidas de probabilidad, entonces el conjunto $\Pi(\mu, \nu)$ no es vacío.*

Demostración. Dadas $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$ podemos construir la medida producto $\mu \otimes \nu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ definida como $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ para todos los bolerianos $A \times B \subset \mathcal{X} \times \mathcal{Y}$. Tenemos que probar que las marginales son μ y ν. Sea $Pr_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ la proyección en la primera componente, luego

$$\begin{aligned}\Big(Pr_{\mathcal{X}\sharp}\mu \otimes \nu\Big)(A \times B) &= \mu \otimes \nu\Big(Pr_{\mathcal{X}}^{-1}(A \times B)\Big) \\ &= (\mu \otimes \nu)(A \times Y) \\ &= \mu(A)\nu(Y) \\ &= \mu(A),\end{aligned}$$

ya que $\nu(\mathcal{Y}) = 1$ por ser ν una medida de probabilidad.

De forma análoga, si $Pr_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ es la proyección en la segunda componente, tenemos que $\Big(Pr_{\mathcal{Y}\sharp}\mu \otimes \nu\Big)(A \times B) = \nu(B)$. Probamos entonces que $Pr_{\mathcal{X}\sharp}\mu \otimes \nu = \mu$ y $Pr_{\mathcal{Y}\sharp}\mu \otimes \nu = \nu$, es decir, $\mu \otimes \nu$ pertenece al conjunto $\Pi(\mu, \nu)$. □

Proposición 2.13. *Si μ y ν son dos medidas de probabilidad entonces $\Pi(\mu, \nu)$ es un conjunto convexo.*

Demostración. Como $\Pi(\mu,\nu) \neq \emptyset$ sabemos que existe al menos un plan de transporte γ_1. Si es el único, no hay nada que probar ya que un conjunto con un único elemento siempre es convexo. Supongamos entonces que existe otro plan de transporte $\gamma_2 \in \Pi(\mu,\nu)$. Para $t \in [0,1]$ definimos $\Gamma_t = t\gamma_1 + (1-t)\gamma_2$. Luego, tenemos que para todo $A \subset \mathcal{X}$ y $B \subset \mathcal{Y}$ boreleanos se cumple

$$\begin{aligned} Pr_{\mathcal{X}\sharp}\Gamma_t(A) &= \Gamma_t\Big(Pr_{\mathcal{X}}^{-1}(A)\Big) \\ &= \Gamma_t(A \times \mathcal{Y}) \\ &= t\gamma_1(A \times \mathcal{Y}) + (1-t)\gamma_2(A \times \mathcal{Y}) \\ &= t\mu(A) + (1-t)\mu(A) \\ &= \mu(A), \end{aligned}$$

y de manera análoga

$$\begin{aligned} Pr_{\mathcal{Y}\sharp}\Gamma_t(B) &= \Gamma_t\Big(Pr_{\mathcal{Y}}^{-1}(B)\Big) \\ &= \Gamma_t(\mathcal{X} \times B) \\ &= t\gamma_1(\mathcal{X} \times B) + (1-t)\gamma_2(\mathcal{X} \times B) \\ &= t\nu(B) + (1-t)\nu(B) \\ &= \nu(B). \end{aligned}$$

Como lo anterior no depende de t tenemos que $t\gamma_1 + (1-t)\gamma_2 \in \Pi(\mu,\nu)$ para todo $t \in [0,1]$, lo cual prueba que el conjunto $\Pi(\mu,\nu)$ es convexo. □

A continuación se presenta y se prueba el teorema fundamental que establece la existencia de un mínimo en el problema de Kantorovich. Para ello, utilizaremos las propiedades de que el conjunto $\Pi(\mu,\nu)$ es convexo y no vacío, así como que $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ es débilmente compacto (ver observación 1.20).

Teorema 2.14. *El problema de Kantorovich siempre admite una solución.*

Demostración. Como $\Pi(\mu,\nu)$ es no vacío, consideramos una sucesión de planes de transportes $\{\gamma_k\}_k \subset \Pi(\mu,\nu)$. Como $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ es débilmente compacto, sabemos que la sucesión $\{\gamma_k\}$ tiene una subsucesión convergente $\{\gamma_{k_l}\}$ en $\Pi(\mu,\nu)$. Queremos probar que el límite $\gamma = \lim_{k_l} \gamma_{k_l} \in \Pi(\mu,\nu)$. Como cada γ_k pertenece

a $\Pi(\mu, \nu)$ tenemos que

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) y d\gamma_k(x, y) = \int_{\mathcal{X}} \varphi(x) d\mu(x) \forall k,$$

por lo tanto

$$\lim_k \int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) y d\gamma_k(x, y) = \int_{\mathcal{X}} \varphi(x) d\mu(x),$$

es decir, $Pr_{\mathcal{X}\sharp}\gamma = \mu$. De forma análoga se prueba que $Pr_{\mathcal{Y}\sharp}\gamma = \nu$, lo cual prueba que $\gamma = \lim_k \gamma_k \in \Pi(\mu, \nu)$, de donde tenemos que $\Pi(\mu, \nu)$ es débilmente cerrado en $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Cómo este último es débilmente compacto tenemos que $\Pi(\mu, \nu)$ también lo es.

Luego, el funcional que es minimizado en el problema de Kantorovich, $\gamma \mapsto \langle c, \gamma \rangle$, es débilmente continuo y por lo tanto admite un mínimo. □

Es fundamental destacar que, una vez establecidas las medidas de probabilidad μ y ν, las soluciones de transporte óptimo derivadas de los problemas de Monge y Kantorovich están considerablemente condicionadas por la elección de la función de costo empleada. Veamos a continuación un ejemplo que ilustra esto. Las imágenes las realizamos en Python y el código está disponible en en GitHub (ver el capítulo 5). Consideremos el siguiente par de medidas discretas:

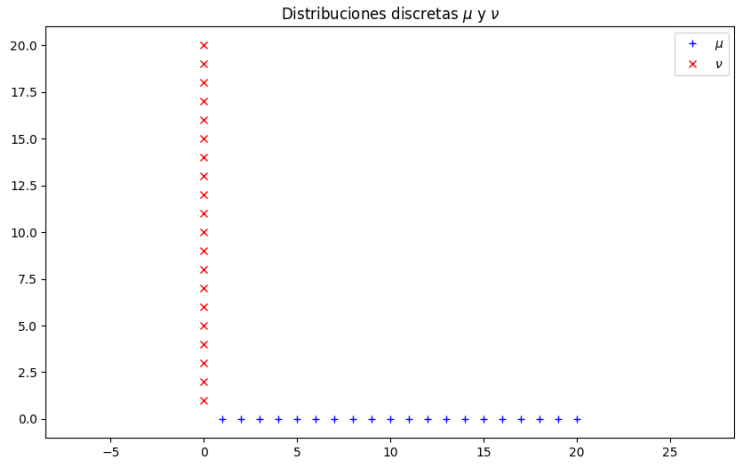


Figura 2.11: En azul la distribución μ, en rojo la distribución ν.

La medida ν está soportada en los puntos de la forma $\{(0, k) \ \ |1 \leq k \leq 20\}$ mientras que la medida μ tiene como soporte los puntos de la forma $\{(i, i \times (-0.001) \ \ | \ \ 1 \leq i \leq 20\}$. Ambas medidas tienen pesos uniformes $\frac{1}{20}$ en cada punto. Multiplicamos la segunda coordenada de los puntos de μ por

-0.001 para que el conjunto resultante sea estrictamente convexo. La figura 2.11 muestra la disposición de los puntos.

A continuación calculamos la matriz de costo utilizando tres normas diferentes: $\|\cdot\|_2$, $\|\cdot\|_2^2$ y $\|\cdot\|_1$. En la figura 2.12 se muestran las respectivas matrices de costo. Los colores más azules representan menor costo mientras que lo más rojos representan mayor costo, así por ejemplo, transportar una unidad de masa de la posición $(0, 1)$ a la posición $(1, 0)$ tiene un costo bajo, mientras que transportar una unidad de masa de la posición $(0, 1)$ a la posición $(17, 0)$ tiene un costo más alto.

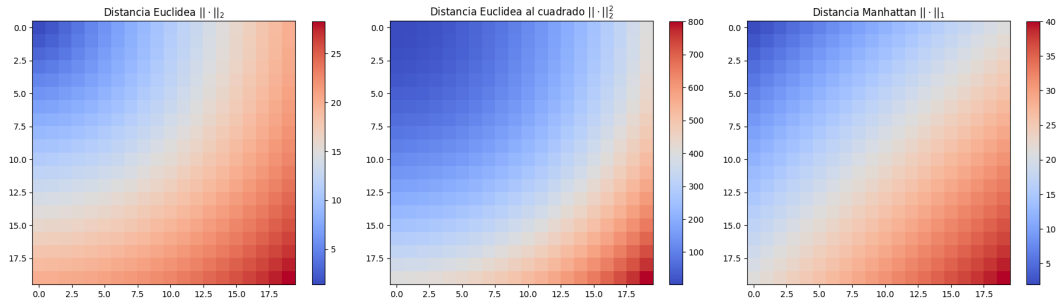


Figura 2.12: Matrices de costo para 3 distancias normas: euclídea $\|\cdot\|_2$, euclídea al cuadrado $\|\cdot\|_2^2$ y Manhattan $\|\cdot\|_1$. Los colores más azules representan un menor costo mientras que los colores más rojos representan un mayor costo.

Como era de esperar, el cuadrado de la distancia euclídea penaliza con mayor fuerza las distancias más grandes, es por esto que la matriz de costo tiene zonas 'menos costosas' y zonas que lo son más, en comparación con la obtenida al utilizar la distancia euclídea. Fijadas las medidas de probabilidad μ, ν y las funciones de costo podemos computar el plan de transporte óptimo. La figura 2.13 muestra las soluciones al problema de transporte óptimo de Monge utilizando las diferentes normas como función de costo.

Cada gráfico representa cómo se mapean los puntos de origen (marcados en azul) a los puntos de destino (marcados en rojo) bajo distintas métricas de distancia. En el gráfico de la izquierda, se utiliza la norma euclídea $||\cdot||_2$ para calcular el costo del transporte. Las líneas rectas indican que los puntos se trasladan directamente en línea recta desde el origen hasta el destino, minimizando la distancia euclidiana. Este tipo de mapeo es intuitivo ya que sigue el camino más corto en un espacio cartesiano regular.

El gráfico central emplea la norma euclídea al cuadrado $||\cdot||_2^2$, que amplifica las distancias mayores más que la norma $||\cdot||_2$. Las trayectorias observadas

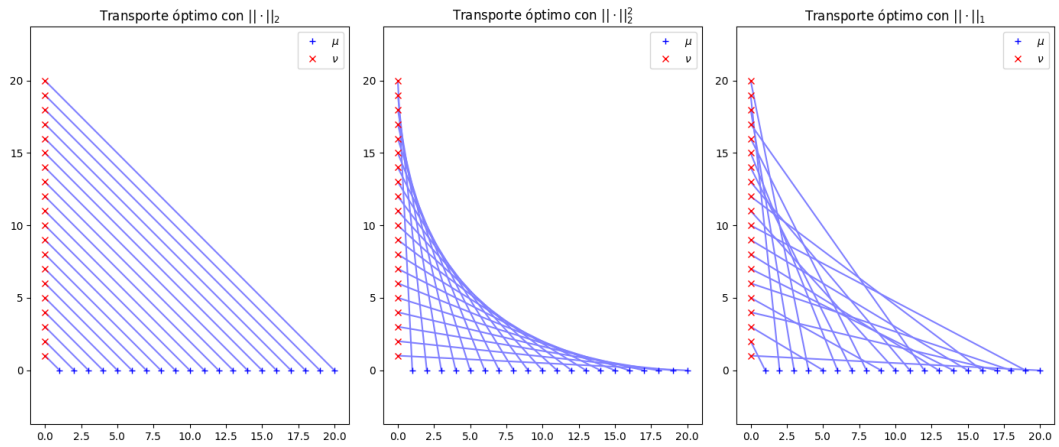


Figura 2.13: Plan de trasporte óptimo entre μ y ν para 3 distancias normas: euclídea $\|\cdot\|_2$, euclídea al cuadrado $\|\cdot\|_2^2$ y Manhattan $\|\cdot\|_1$. Es importante notar que al usar la distancia euclídea (izquierda) las lineas no se cruzan.

muestran cómo esta norma penaliza más fuertemente las distancias largas, resultando en un mapeo que no es lineal y que refleja una distribución diferente de las masas para minimizar el costo cuadrático. Esto provoca trayectorias que equilibran los desplazamientos en cada eje para evitar excesos en uno solo, resultando en caminos más directos en términos de distancia máxima.

Por último, en el gráfico de la derecha, se utiliza la norma Manhattan $||\cdot||_1$, que mide la distancia como la suma de las diferencias absolutas en cada dimensión. Las líneas reflejan trayectorias que siguen rutas más complejas, minimizando la distancia total en términos de desplazamientos horizontales y verticales combinados. En resumen, la elección de la norma afecta significativamente las trayectorias de transporte óptimo.

2.1.3. Geometría del transporte óptimo en el plano

Presentamos a continuación un resultado interesante cuando trabajamos en $\mathcal{X} = \mathbb{R}^2 = \mathcal{Y}$ y cuando μ y ν son medidas discretas. Si usamos como función de costo la distancia euclídea, Monge probó que los segmentos de recta que representan el transporte óptimo no se pueden cruzar. La figura 2.14 muestra la situación.

El resultado es realidad más amplio, y vale para $\mathbb{R}^n$. La formulación completa así como su demostración puede encontrarse en capítulo 5 del libro de Villani (2009), bajo el nombre de "cyclical monotonicity". Presentamos a con-

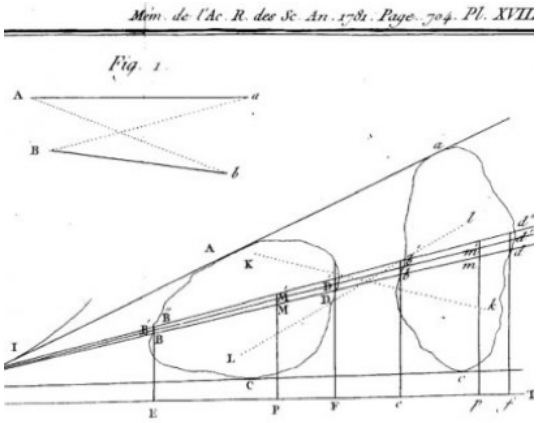


Figura 2.14: Al utilizar como función de costo la distancia euclídea los segmentos que representan el transporte óptimo no se pueden cruzar. Figura tomada del trabajo original de Monge (1781).

tinuación el caso en $\mathbb{R}^2$, el cual es un resultado propio y puede ser probado directamente con la teoría introducida hasta este momento.

Proposición 2.15. *Sean $\mathcal{X} = \{x_1, \dots, x_n\}$ un conjunto finito en $\mathbb{R}^2$ contenidos en una recta y $\mu = \sum_{i=1}^n \frac{1}{n}\delta_{x_i}$ la medida uniforme sobre $\mathcal{X}$. Supongamos que $\mathcal{Y} = \{y_1, \dots, y_n\}$ se obtiene a partir $\mathcal{X}$ rotando cierto ángulo θ y $\nu = \sum_{i=1}^n \frac{1}{n}\delta_{y_i}$ es la medida uniforme sobre $\mathcal{Y}$. Si utilizamos como función de costo c la distancia euclídea, entonces la asignación resultante del transporte óptimo entre μ y ν coincide con la rotación.*

Demostración. Para comenzar notamos que en el contexto donde $\mathcal{X} = \{x_1, \dots, x_n\}$ e $\mathcal{Y} = \{y_1, \dots, y_n\}$ son dos conjuntos finitos con la misma cantidad de puntos y las medidas μ y ν son uniformes entonces cualquier transporte es una permutación del conjunto de índices $\{1, \dots, n\}$, es decir, $\{T : \mathcal{X} \to \mathcal{Y} \mid T_\sharp \mu = \nu\} = \{\sigma : \{1, \dots, n\} \to \{1, \dots, n\} \mid \sigma \text{ es una biyección}\}$, donde identificamos $T_\sigma(x_i) = y_{\sigma(i)}$. En este contexto, la rotación se identifica con la permutación identidad, es decir, $T(x_i) = y_i$.

La idea consiste en demostrar que, al desarmar cada cruce, el costo disminuye. Además, desarmar todos los cruces nos lleva a la permutación identidad, alcanzando así nuestro objetivo. Empecemos por formalizar el concepto de "desarmar un cruce".

Supongamos que tenemos 4 índices $i, j, k, l \in \{1, \dots, n\}$ tales que $T(x_i) =$

y_j y $T(x_k) = y_l$. Diremos que hay un cruce si los segmentos $\overline{x_i T(x_i)}$ y $\overline{x_k T(x_k)}$ se intercepten. La figura 2.15 bosqueja esta situación.

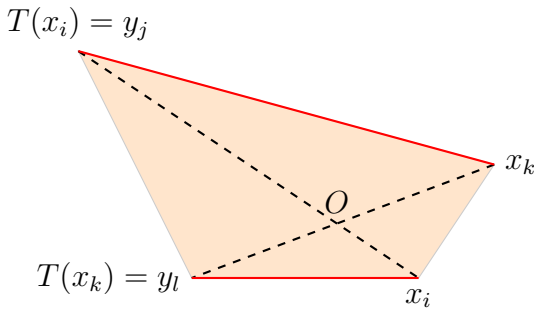


Figura 2.15: Los segmentos punteados representan el transporte $T(x_i) = y_j$ y $T(x_k) = y_l$, los cuales se cruzan en el punto O. Los segmentos rojos representa el transporte obtenido al desarmar el cruce.

Cuando decimos "desarmar el cruce", nos referimos a emplear un transporte alternativo $\widetilde{T}$, que es igual a T para todos los índices $\{1, \ldots, n\} \backslash \{i, k\}$ y para $\{i, k\}$ permuta las imágenes de x_i y x_k a través de T, es decir:

$$\widetilde{T}(x_i) = T(x_k), \quad \widetilde{T}(x_k) = T(x_i)$$

Si nos concentramos únicamente en los puntos x_i y x_k, entonces el transporte T está representado por los segmentos punteados negros mientras que el transporte $\widetilde{T}$ está representado por los segmentos rojos de la figura 2.15.

Nuestro objetivo es probar que el costo asociado al transporte $\widetilde{T}$ es estrictamente menor al costo asociado al transporte T. Como estos dos transportes solo difieren en x_i y x_k, probar nuestro objetivo es equivalente a probar que la suma de los lados rojos del polígono convexo de la figura 2.15 es estrictamente menor que la suma de sus diagonales.

Como estamos utilizando como función de costo la distancia euclídea, sabemos que vale la desigualdad triangular, a partir de estas tenemos que $\overline{y_j O} + \overline{O x_k} > \overline{y_j x_k}$ y además $\overline{y_l O} + \overline{O x_i} > \overline{y_l x_i}$, sumando ambas desigualdades llegamos a

$$\underbrace{\overline{y_j O} + \overline{O x_i}}_{\overline{y_j x_i}} + \underbrace{\overline{y_l O} + \overline{O x_k}}_{\overline{y_l x_k}} > \overline{y_j x_k} + \overline{y_l x_i}$$

□

El resultado anterior nos será útil en las simulaciones, donde mostraremos

como utilizar el transporte óptimo para lograr una adaptación de dominio de una regresión lineal a partir de una rotación.

2.1.4. Distancia de Wasserstein

Un concepto importante utilizado en el transporte óptimo es la definición de la conocida distancia de Wasserstein, una herramienta fundamental para medir la diferencia entre dos distribuciones de probabilidad. Esta distancia, originada en los trabajos de Leonid Kantorovich y Maurice Fréchet (1939), ha adquirido un papel central en numerosas aplicaciones prácticas y teóricas. Es especialmente relevante en campos como la teoría de probabilidad, la estadística, el aprendizaje automático y, más recientemente, en el análisis de datos de alta dimensión.

Definición 2.16 (Distancia de Wasserstein). *Sean $(\mathcal{X}, d)$ un espacio métrico compacto, $\mu, \nu \in \mathcal{P}(\mathcal{X})$ dos medidas de probabilidad en $\mathcal{X}$ y $p \in [1, \infty)$. Definimos la distancia de Wasserstein de orden p entre μ y ν como*

$$W_p(\mu, \nu) = \Big(\inf_{\gamma \in \Pi(\mu,\nu)} \Big\{ \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\gamma(x,y) \Big\} \Big)^{\frac{1}{p}},$$

donde $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \ \mid \ \gamma$ tiene medidas marginales μ y $\nu\}$.

Un caso muy particular es cuando p vale 1, ya que en ese caso tenemos que W_1 es el costo de Kantorovich obtenido al utilizar cualquier solución óptima γ^*.

Observación 2.17. *En la literatura puede encontrase la distancia $W_1(\mu, \nu)$ como 'Earth Mover's Distance', dado su origen en el problema de transporte óptimo y la interpretación de Monge del mismo.*

A partir de la formulación probabilística, podemos interpretar la distancia $W_p(\mu, \nu)$ como

$$W_p(\mu, \nu) = \Big(\inf_{\gamma} \mathbb{E}_{(X,Y)\sim\gamma} \Big\{ d(X,Y)^p \Big\} \Big)^{\frac{1}{p}}$$

Veamos a continuación dos ejemplos para ganar un poco de intuición: la distancia de Wasserstein entre dos deltas de Dirac y la distancia de Wasserstein entre dos gaussianas univariadas.

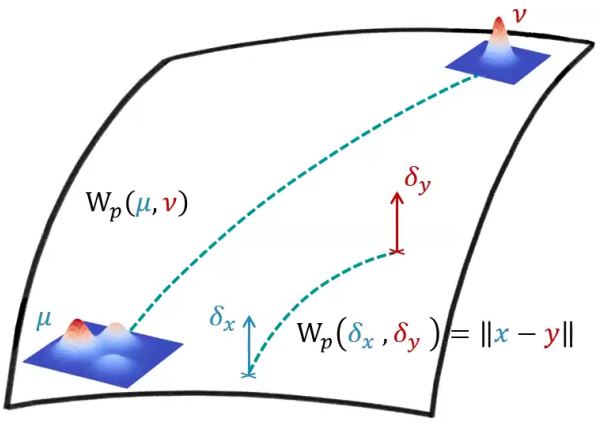


Figura 2.16: Distancia de Wasserstein entre dos medidas arbitrarias y entre dos deltas de Dirac. Figura autoria de Ziv Goldfeld.

Ejemplo 2.18. *Sean x, y dos puntos en $\mathbb{R}^n$, definimos $\mu = \delta_x$ y $\nu = \delta_y$. Si el costo $c(x, y)$ viene dado por la distancia euclídea $c(x, y) = ||x - y||_2$ entonces*

$$W_p(\delta_x, \delta_y) = ||x - y||_2.$$

En efecto, el único transporte (tanto de Monge como de Kantorovich) posible es $T(x) = y$, de donde,

$$W_p(\delta_x, \delta_y) = \Big(\int_{\mathbb{R}^n \times \mathbb{R}^n} ||x - y||_2^p d\delta_x d\delta_y \Big)^{\frac{1}{p}} = \Big(||x - y||_2^p \Big)^{\frac{1}{p}} = ||x - y||_2.$$

El ejemplo anterior nos dice que la distancia entre los puntos x e y en el espacio base $\mathbb{R}^n$ es igual a la distancia de Wasserstein de las medidas δ_x y δ_y. La figura 2.16 muestra un bosquejo de la situación. Veamos a continuación la distancia de Wasserstein entre dos gaussianas univariadas.

Ejemplo 2.19 (Distancia de Wasserstein entre dos Gaussianas univariadas). *Sean $X \sim \mu = \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \nu = \mathcal{N}(\mu_2, \sigma_2^2)$ dos gaussianas en $\mathbb{R}$. Afirmamos que el mapa $T : \mathbb{R} \to \mathbb{R}$ definido como*

$$T(X) = \frac{\sigma_2}{\sigma_1}(X - \mu_1) + \mu_2,$$

es un transporte entre μ y ν.

En efecto tenemos que

$$\mathbb{E}T(X) = \frac{\sigma_2}{\sigma_1}\Big(\underbrace{\mathbb{E}X - \mu_1}_{0}\Big) + \mu_2 = \mu_2,$$

y por otro lado,

$$varT(X) = \frac{\sigma_2^2}{\sigma_1^2} varX = \frac{\sigma_2^2}{\sigma_1^2}\sigma_1^2 = \sigma_2^2.$$

Por último como X está normalmente distribuida y T es una transformación afín entonces $T(X)$ es una distribución normal, que por lo anterior, tiene media y varianza μ_2 y σ_2^2 respectivamente, es decir, $T(X) = Y$.

Como este mapa es la derivada de la función convexa

$$\varphi(x) = \frac{\sigma_2}{2\sigma_1}(x - \mu_1)^2 + \mu_2 x,$$

por el teorema de Brenier (ver más adelante teorema 2.25) sabemos que si utilizamos el costo $c(x, y) = ||x-y||_2^2$ entonces T es el único plan de transporte óptimo. La distancia de Wasserstein para $p = 2$ queda

$$W_2^2(\mu, \nu) = \int_{\mathbb{R}} \Big(\frac{\sigma_2}{\sigma_1}(x - \mu_1) + \mu_2 - x\Big) d\mu(x) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2.$$

En el caso particular donde μ y ν tienen misma varianza, tenemos que

$$W_2^2(\mu, \nu) = (\mu_1 - \mu_2)^2 \quad \textit{de donde} \quad W_2(\mu, \nu) = |\mu_1 - \mu_2|,$$

es decir que la distancia de Wasserstein entre dos gaussianas univariadas con misma varianza es simplemente la distancia euclídea entre sus medias, en particular es independiente de que tan concentrada (σ pequeño) o dispersa (σ grande) sean estas medidas. La figura 2.17 muestra con dos ejemplos como la distancia $W_2(\mu, \nu)$ no cambia al modificar la varianza.

Cuando trabajamos con dos medidas guassianas en $\mathbb{R}^n$, digamos $\mu \sim \mathcal{N}(\mu_1, \Sigma_1)$ y $\nu \sim \mathcal{N}(\mu_2, \Sigma_2)$ entonces el transporte óptimo es

$$T(x) = \mu_2 + A(x - \mu_1),$$

donde A es una matriz simétrica tal que $A\Sigma_1 A = \Sigma_2$. En este contexto, la

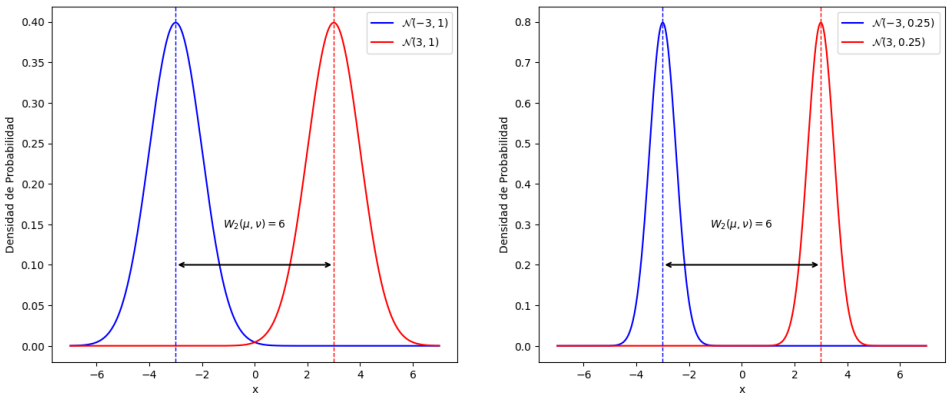


Figura 2.17: Distancia de Wasserstein entre gaussianas univariadas con $p = 2$ y $c(x, y) = |x - y|^2$. Si tienen misma varianza entonces $W_2(\mu, \nu)$ solo depende de la distancia entre sus medias.

distancia de Wasserstein queda

$$W_2^2(\mu,\nu) = ||\mu_1 - \mu_2||^2 + tr\Big(\Sigma_1 + \Sigma_2 - 2\Big(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\Big)^{\frac{1}{2}})\Big).$$

Una prueba de este resultado puede encontrase en el libro de Gabriel Peyré y Marco Cuturi (2019).

Veamos a continuación como es la distancia de Wasserstein entre dos medidas con densidad en $\mathbb{R}$ arbitrarias. Recordamos primero la proposición 1.7:

Si μ es una medida de probabilidad en $\mathbb{R}$ y $\mathcal{U}$ es la distribución uniforme en $[0, 1]$ entonces

$$(\mathcal{C}_\mu)^{-1}_\sharp \mathcal{U} = \mu,$$

y por lo tanto

$$(\mathcal{C}_\mu)_\sharp \mu = \mathcal{U}.$$

En particular, lo anterior implica que si μ y ν son medidas de probabilidad en $\mathbb{R}$ entonces el mapa $T = \mathcal{C}_\nu^{-1} \circ \mathcal{C}_\mu$ satisface que $T_\sharp \mu = \nu$, es decir, es un transporte entre μ y ν. Se puede probar además, utilizando el teorema de Brenier (ver más adelante teorema 2.25), que este transporte es óptimo, ver por ejemplo en el libro de Gabriel Peyré y Marco Cuturi (Peyre y Cuturi, 2019).

Luego, si utilizamos como costo la distancia $c(x, y) = |x - y|^2$, tenemos

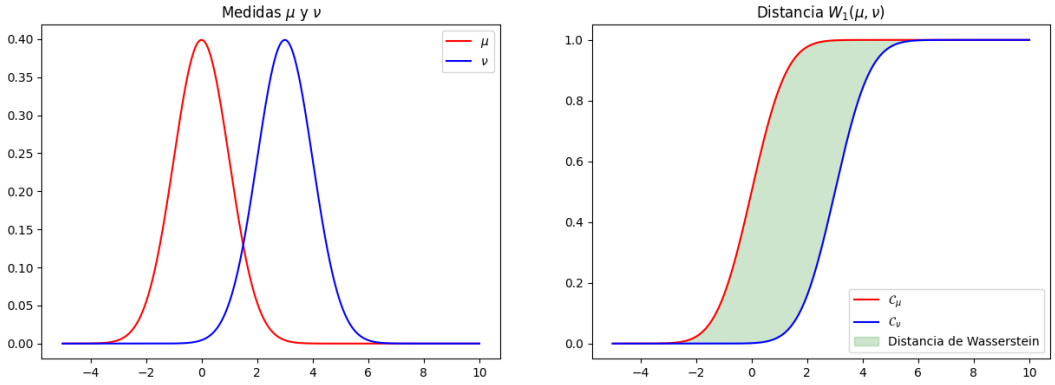


Figura 2.18: Distancia de Wasserstein entre dos medidas de probabilidad en $\mathbb{R}$ utilizando el costo $c(x, y) = |x - y|^2$ y $p = 1$. A la izquierda las medidasμ y ν, a la derecha sus funciones acumulativas $\mathcal{C}_\alpha$ y $\mathcal{C}_\nu$. El área entre las dos funciones acumulativas es igual a la distancia de Wasserstein entre las medidas.

para cualquier $p \geq 1$ que la distancia de Wasserstein es

$$\begin{aligned} W_p(\mu, \nu)^p &= \int_{\mathbb{R}} \left| x - \mathcal{C}_\nu^{-1}\Big(\mathcal{C}_\mu(x)\Big) \right|^p d\mu(x) \\ &= \int_0^1 \left| \mathcal{C}_\mu^{-1}(r) - \mathcal{C}_\nu^{-1}(r) \right|^p dr \\ &= ||\mathcal{C}_\mu^{-1} - \mathcal{C}_\nu^{-1}||_{L^p([0,1])}. \end{aligned}$$

Esta igualdad nos dice que a través del mapa $\mu \mapsto \mathcal{C}_\mu^{-1}$, la distancia de Wasserstein es isométrica a un espacio lineal con la norma L^p.

El caso $p = 1$ es particular, por qué obtenemos que la distancia $W_1(\mu, \nu)$ es al área comprendida entre las funciones acumulativas de μ y ν, explícitamente

$$W_1(\mu, \nu) = ||\mathcal{C}_\mu - \mathcal{C}_\nu||_{L^1(\mathbb{R})} = \int_{\mathbb{R}} \Big| \mathcal{C}_\mu(x) - \mathcal{C}_\nu(x) \Big| dx = \int_{\mathbb{R}} \Big| \int_{-\infty}^{x} d(\mu - \nu) \Big| dx.$$

La figura 2.18 muestra un ejemplo con dos gaussianas.

Probaremos a continuación que bajos ciertas circunstancias, la 'distancia de Wasserstein' es efectivamente una distancia.

Teorema 2.20. *En el caso donde $\mathcal{X} = \mathcal{Y}$ es un espacio métrico compacto con una distancia d entonces*

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{X}} d(x, y)^p d\pi(x, y) \right\} \right)^{\frac{1}{p}}$$

es una distancia en el espacio de las medidas $\mathcal{M}^+(\mathcal{X})$.

La siguiente demostración es una adaptación de la versión presentada en el libro de Villani (2009).

Demostración. Como d es una distancia es claro que $d(x,y) = d(y,x)$ por lo que $W_p(\mu,\nu) = W_p(\nu,\mu)$, y por lo tanto, es simétrica. $W_p(\mu,\mu) = 0$ ya que en dicho caso $d(x,y)$ es identicamente nulo. La no negatividad es trivial.

Para probar que $W_p(\mu,\nu) = 0$ implica que $\mu = \nu$ observamos que si $W_p(\mu,\nu) = 0$ entonces necesariamente la medida π está soportada en la recta $x = y$ (ya que el costo $c(x,y)$ tiene que ser 0 para todo par (x,y)). Lo anterior implica que para cualquier función continua φ se cumple que

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) d\pi(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} \varphi(y) d\pi(x,y),$$

de donde

$$\begin{aligned}
\int_{\mathcal{X}} \varphi(x) d\mu(x) &= \int_{\mathcal{X}} \varphi(x) \int_{\mathcal{Y}} d\pi(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) d\pi(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \varphi(y) d\pi(x,y) \\
&= \int_{\mathcal{Y}} \varphi(y) \int_{\mathcal{X}} d\pi(x,y) \\
&= \int_{\mathcal{Y}} \varphi(y) d\nu(y).
\end{aligned}$$

Como lo anterior vale para cualquier función medible φ tenemos que $\mu = \nu$.

Nos resta probar la desigualdad triangular, para esto utilizaremos el gluing lemma. Para esto consideramos las medidas $\mu_1, \mu_2, \mu_3 \in \mathcal{M}^+(\mathcal{X})$ y los planes de transporte óptimos $\gamma_{12} \in \Pi(\mu_1,\mu_2)$ y $\gamma_{23} \in \Pi(\mu_2,\mu_3)$. Por el gluing lemma 1.25 tenemos que existe una medida $\gamma \in \mathcal{M}^+(\mathcal{X}\times\mathcal{X}\times X)$ cuyas marginales son γ_{12} y γ_{23}, es decir, $\gamma(A\times\mathcal{X}) = \gamma_{12}(A)$ y $\gamma(\mathcal{X}\times B) = \gamma_{23}(B)$ para todo boreliano $(A,B) \in \mathcal{X}\times\mathcal{X}$.

Definimos la medida $\pi \in \mathcal{M}^+(\mathcal{X}\times\mathcal{X})$ como la proyección de γ en la primera y tercera coordenada, es decir,

$$\pi(C) = \gamma\Big(\Big\{(x,y,z) \in \mathcal{X}\times\mathcal{X}\times\mathcal{X} \;\;|\;\; (x,z) \in C\Big\}\Big),$$

para todo conjunto medible $C \in \mathcal{X} \times \mathcal{X}$. A continuación probamos que $\pi \in \Pi(\mu_1, \mu_3)$:

$$\pi(A \times \mathcal{X}) = \gamma\Big(\Big\{(x, y, z) \in \mathcal{X} \times \mathcal{X} \times \mathcal{X} \;\; | \;\; x \in A\Big\}\Big) = \gamma_{12}(A \times \mathcal{X}) = \mu_1(A),$$

de forma análoga

$$\pi(\mathcal{X} \times B) = \gamma\Big(\Big\{(x, y, z) \in \mathcal{X} \times \mathcal{X} \times \mathcal{X} \;\; | \;\; z \in B\Big\}\Big) = \gamma_{23}(\mathcal{X} \times B) = \mu_3(B),$$

por lo tanto $\pi \in \Pi(\mu_1, \mu_3)$.

Además por ser $d(\cdot, \cdot)$ una distancia tenemos que

$$\Big(\int_{\mathcal{X}\times\mathcal{X}} d(x, z)^p d\pi(x, z)\Big)^{\frac{1}{p}} \leq \Big(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} (d(x, y) + d(y, z))^p d\gamma(x, y, z)\Big)^{\frac{1}{p}}.$$

Por otro lado, utilizando la desigualdad de Minkowski tenemos

$$\begin{aligned}\Big(\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{X}} (d(x, y) + d(y, z))^p d\gamma(x, y, z)\Big)^{\frac{1}{p}} &\leq \Big(\int_{\mathcal{X}\times\mathcal{X}} d(x, y)^p d\gamma(x, y, z)\Big)^{\frac{1}{p}} \\ &\quad + \Big(\int_{\mathcal{X}\times\mathcal{X}} d(y, z)^p d\gamma(x, y, z)\Big)^{\frac{1}{p}}.\end{aligned}$$

En la primera integral de lado derecho podemos cambiar $d\gamma(x, y, z)$ por $d\gamma_{12}(x, y)$ dado que la variable z no aparece en el integrando. De manera similar, en la segunda integral de la derecha, podemos cambiar $d\gamma(x, y, z)$ por $d\gamma_{23}(y, z)$ ya que no aparece la variable x en el integrando. Finalmente, utilizando que π es una medida "acoplada" sobre γ_{12} y γ_{23} podemos escribir

$$\begin{aligned}\Big(\int_{\mathcal{X}\times\mathcal{X}} d(x, z)^p d\pi(x, z)\Big)^{\frac{1}{p}} &\leq \Big(\int_{\mathcal{X}\times\mathcal{X}} d(x, y)^p d\pi_1(x, y)\Big)^{\frac{1}{p}} \\ &\quad + \Big(\int_{\mathcal{X}\times\mathcal{X}} d(y, z)^p d\pi_2(y, z)\Big)^{\frac{1}{p}},\end{aligned}$$

lo cual es equivalente a $W_p(\mu_1, \mu_3) \leq W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3)$. □

Interpolación de medidas

Una aplicación útil de la distancia de Wasserstein es la interpolación de medidas. Usualmente la interpolación entre dos medidas μ_0 y μ_1 en un espacio

métrico $\mathcal{X}$ se realizaba mediante lo que se denomina interpolación ℓ_2:

$$\mu_t = (1-t)\mu_0 + t\mu_1, \quad t \in [0,1].$$

Este método, si bien es sencillo, conlleva a que en cada tiempo $t \in (0,1)$ se tiene una mezcla de μ_0 y μ_1, así por ejemplo aunque μ_0 y μ_1 sean distribuciones normales tendremos que μ_t es una mezcla de normales. Por otro lado, podemos definir la interpolación utilizando la distancia de Wasserstein de la siguiente manera:

Definición 2.21 (Interpolación de Wasserstein). *Sean μ_0 y μ_1 dos medidas en un espacio métrico compacto $\mathcal{X}$. La interpolación de Wasserstein entre μ_0 y μ_1 es*

$$\mu_t = \arg\min_{\mu} \left\{ (1-t)W_2(\mu,\mu_0)^2 + tW_2(\mu,\mu_1)^2 \right\},$$

donde $t \in [0,1]$.

Si bien esta formulación es más abstracta y costosa computacionalmente (debido a que para cada t hay que resolver un problema de optimización), resulta que la medida μ_t preserva mejor la estructura de las medidas originales. Siguiendo con el ejemplo donde μ_0 y μ_1 son dos gaussianas, se tiene que μ_t es también una gaussiana para cada $t \in [0,1]$. La figura 2.19 muestra como se diferencia μ_t al utilizar la distancia ℓ_2 o la distancia de Wasserstein.

Otra aplicación reciente e interesante surgió de la distancia de Wasserstein en algunos trabajos de modelos generativos, en particular el artículo de Arjovsky, Chintala y Bottou (2017) utiliza la distancia de Wasserstein para comparar distribuciones de probabilidad al momento de entrenar una GAN ('Generative Adversarial Networks').

Ahora que introdujimos la distancia de Wasserstein mostraremos con algunos ejemplos las diferencias entre ella y la divergencia de Kullback-Leibler también utilizada para medir discrepancias entre medidas. Comencemos con un ejemplo discreto.

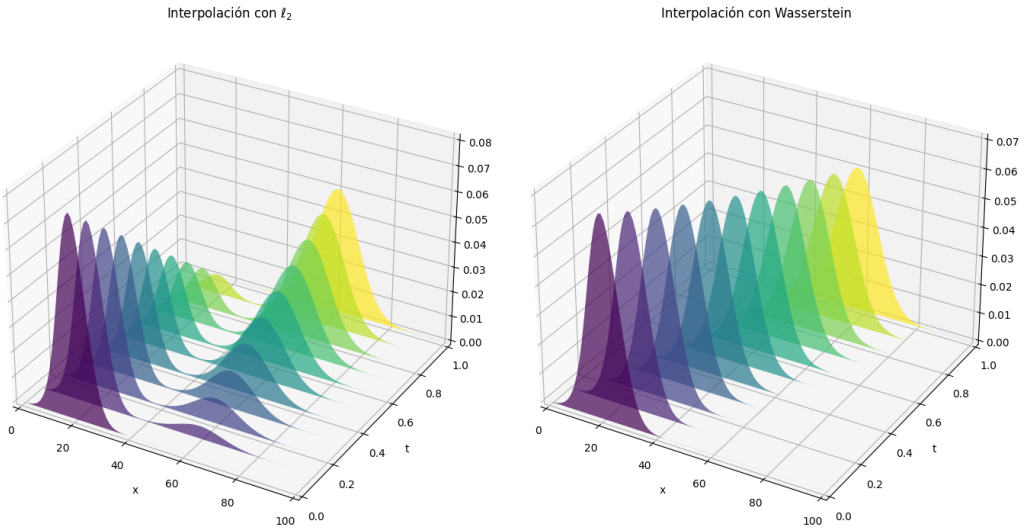


Figura 2.19: Interpolación entre dos gaussianas, a la izquierda utilizando la distancia ℓ_2, a la derecha utilizando la distancia de Wasserstein. Las gaussianas violeta y amarilla son μ_0 y μ_1 respectivamente. Al utilizar la distancia ℓ_2 se tiene que μ_t es una mezcla de gaussianas para $t \in (0,1)$ mientras que al utilizar la distancia de Wasserstein se tiene que μ_t es una gaussiana para todo $t \in [0,1]$.

Ejemplo 2.22. *Supongamos que tenemos las siguientes medidas discretas en* $\mathbb{R}$*:*

$$\begin{aligned}
\mu &= \frac{6}{10}\delta_0 + \frac{1}{10}\delta_1 + \frac{1}{10}\delta_2 + \frac{1}{10}\delta_3 + \frac{1}{10}\delta_4, \\
\nu_1 &= \frac{1}{10}\delta_0 + \frac{1}{10}\delta_1 + \frac{1}{10}\delta_2 + \frac{1}{10}\delta_3 + \frac{6}{10}\delta_4, \\
\nu_2 &= \frac{1}{10}\delta_0 + \frac{6}{10}\delta_1 + \frac{1}{10}\delta_2 + \frac{1}{10}\delta_3 + \frac{1}{10}\delta_4.
\end{aligned}$$

La figura 2.20 muestra estas medidas como histogramas. Comencemos calculando la divergencia de Kullback-Leibler entre μ *y* ν_1 *y entre* μ *y* ν_2*.*

$$\begin{aligned}
D_{KL}(\mu||\nu_1) &= \sum_{k=0}^{4} \mu_k \log\left(\frac{\mu_k}{\nu_{1k}}\right) \\
&= \frac{6}{10}\log(6) + \frac{1}{10}\log\left(1\right) + \frac{1}{10}\log\left(1\right) + \frac{1}{10}\log\left(1\right) + \frac{1}{10}\log\left(\frac{1}{6}\right) \\
&= \frac{6}{10}\log(6) + \frac{1}{10}\log\left(\frac{1}{6}\right)
\end{aligned}$$

Por otro lado,

$$
\begin{aligned}
D_{KL}(\mu||\nu_2) &= \sum_{k=0}^{4} \mu_k \log\left(\frac{\mu_k}{\nu_{1k}}\right) \\
&= \frac{6}{10}\log(6) + \frac{1}{10}\log\left(\frac{1}{6}\right) + \frac{1}{10}\log\left(1\right) + \frac{1}{10}\log\left(1\right) + \frac{1}{10}\log\left(1\right) \\
&= \frac{6}{10}\log(6) + \frac{1}{10}\log\left(\frac{1}{6}\right),
\end{aligned}
$$

y por lo tanto $D_{KL}(\mu||\nu_1) = D_{KL}(\mu||\nu_2)$. *Veamos a continuación qué sucede con la distancia de Wasserstein: utilizando como función de costo la norma euclídea* $||\cdot||_2^2$. *Por estar usando la norma* $||\cdot||_2^2$ *como función de costo, mientras menos masa movamos menos costoso será el transporte. Con esto en mente es sencillo convencerse que la manera más eficiente de mover* μ *a* ν_1 *es*

- *Dejar las masas de las posiciones* $1, 2, 3$ *y* 4 *fijas.*
- *Dejar una unidad de masa de la posición* 0 *fija.*
- *Mover las restantes* 5 *unidades de masa de la posición* 0 *a la posición* 4.

Observar que lo anterior no puede ser una solución al problema de Monge dado que estamos haciendo una separación de masa. Luego la distancia de Wasserstein entre μ *y* ν_1 *es*

$$
W_2(\mu, \nu_1) = \left(5 \times ||\delta_0 - \delta_4||_2^2\right)^{\frac{1}{2}} = \sqrt{5 \times 16} = \sqrt{80}.
$$

De manera similar, el plan de transporte óptimo entre μ *y* ν_2 *es*

- *Dejar las masas de las posiciones* $1, 2, 3$ *y* 4 *fijas.*
- *Dejar una unidad de masa de la posición* 0 *fija.*
- *Mover las restantes* 5 *unidades de masa de la posición* 0 *a la posición* 1.

A partir de esto, la distancia de Wasserstein entre μ *y* ν_2 *es*

$$
W_2(\mu, \nu_2) = \left(5 \times ||\delta_0 - \delta_1||_2^2\right)^{\frac{1}{2}} = \sqrt{5 \times 1} = \sqrt{5},
$$

por lo que $W_2(\mu, \nu_2) < W_2(\mu, \mu_1)$.

El ejemplo anterior muestra una diferencia importante entre la divergencia de Kullback-Leibler y la distancia de Wasserstein: la primera mide diferencias verticales mientras que la segunda mide distancias horizontales. Esto se debe a que en la divergencia Kullback-Leibler aparece un termino de la forma $\frac{d\mu}{d\nu}$,

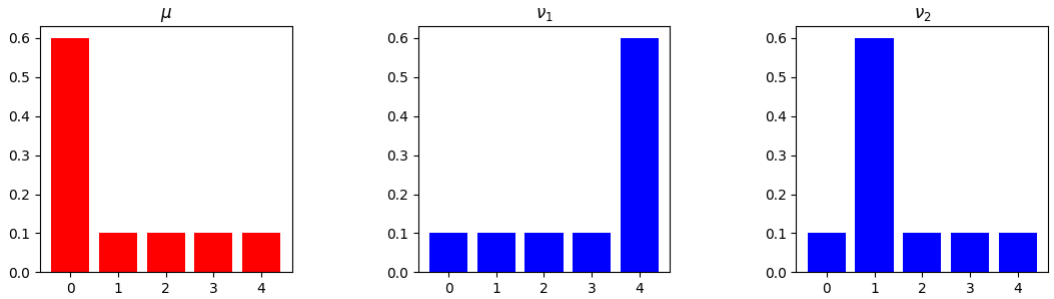


Figura 2.20: De izquierda a derecha: Las medidas μ, ν_1 y ν_2. Se tiene que $D_{KL}(\mu, \nu_1) = D_{KL}(\mu, \nu_2)$ mientras que $W_2(\mu, \nu_1) > W_2(\mu, \nu_2)$

cociente punto a punto entre funciones, lo cual cuantifica diferencias verticales, mientras que en la distancia de Wasserstein aparece la distancia entre los puntos $d(x, y)$, la cual cuantifica diferencias horizontales.

2.2. Relación entre las formulaciones de Monge y Kantorovich

Gaspard Monge y Leonid Kantorovich abordaron el problema de transporte óptimo desde perspectivas significativamente distintas. Monge, en el siglo XVIII, planteó el problema en términos geométricos, buscando un mapa $T : \mathcal{X} \to \mathcal{Y}$ que trasladara directamente una masa de un punto $x \in \mathcal{X}$ a otro punto $y \in \mathcal{Y}$ de la forma más eficiente posible, minimizando el costo del transporte. Su enfoque se centraba en la asignación directa y determinista de masas entre los puntos de origen y destino. Por otro lado, Kantorovich, en el siglo XX, reformuló el problema utilizando métodos de optimización lineal y planteó la búsqueda de una distribución probabilística conjunta γ sobre $\mathcal{X} \times \mathcal{Y}$ que describiera cómo distribuir las cargas de manera óptima entre múltiples destinos. Esta formulación permitió un análisis más flexible transformando el problema en uno de programación lineal, lo que facilitó su resolución y abrió nuevas posibilidades en la teoría de la optimización. Podemos resumir lo visto hasta el momento en la tabla 2.1.

En esta sección mostraremos la relación que hay entre ambas formulaciones así como sus implicancias.

A partir de los ejemplos presentados en las secciones anteriores es claro que hay situaciones donde no existe una solución al problema de Monge pero si

existe una solución para el problema de Kantorovich. Una pregunta directa que se desprende es si hay alguna relación entre las soluciones, de existir, de ambas formulaciones. Comenzamos probando que a partir de un mapa de transporte T (Monge) se puede construir un plan de transporte γ_T (Kantorovich).

Aspecto	**Monge**	**Kantorovich**
¿Qué se busca?	Una función $T : \mathcal{X} \to \mathcal{Y}$	Una medida $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$
Existencia de la solución	No siempre existe	Siempre existe
Unicidad de la solución	No siempre única	No siempre única

Tabla 2.1: Diferencias entre los problemas de Monge y Kantorovich

Proposición 2.23. *Si $T : \mathcal{X} \to \mathcal{Y}$ es un mapa de transporte del problema de Monge entonces $\gamma_T = (id \times T)_\sharp \mu$ es un plan de transporte en el sentido de Kantorovich, y por lo tanto, una solución factible al problema de Kantorovich.*

Demostración. Sea T un transporte entre μ y ν, es decir $T_\sharp \mu = \nu$. Definamos $\gamma_T = (id \times T)_\sharp \mu$, explícitamente, dado $A \times B \subset \mathcal{X} \times \mathcal{Y}$, tenemos

$$\gamma_T(A \times B) = \mu\Big(\{x \in \mathcal{X} \ \mid \ x \in A \text{ y } T(x) \in B\}\Big).$$

Veamos que $\gamma_T \in \Pi(\mu, \nu)$. Por un lado

$$\begin{aligned} Pr_{\mathcal{X}\sharp}\gamma_T(A) &= \gamma_T\Big(Pr_{\mathcal{X}}^{-1}(A)\Big) \\ &= \gamma_T\Big(A \times \mathcal{Y}\Big) \\ &= \mu\Big(\{x \in \mathcal{X} \ \mid \ x \in A, T(x) \in \mathcal{Y}\}\Big) = \mu(A) \end{aligned}$$

dado que $T(x)$ siempre pertenece a $\mathcal{Y}$. Por otro lado,

$$\begin{aligned} Pr_{\mathcal{Y}\sharp}\gamma_T(B) &= \gamma_T\Big(Pr_{\mathcal{Y}}^{-1}(B)\Big) \\ &= \gamma_T\Big(\mathcal{X} \times B\Big) \\ &= \mu\Big(\{x \in \mathcal{X} \ \mid \ x \in \mathcal{X}, T(x) \in B\}\Big) \\ &= \mu\Big(T^{-1}(B)\Big) = T_\sharp\mu(B) = \nu(B) \end{aligned}$$

donde en la última igualdad usamos el hecho de que T es un transporte entre μ y ν. □

A continuación, analizaremos una proposición que sostiene que la solución obtenida al resolver el problema de Kantorovich es, en términos de eficiencia, superior o, en el peor de los casos, igual a la solución obtenida por el problema de Monge.

Proposición 2.24. *Si γ^* y T^* son soluciones óptimas a los problemas de Kantorovich y Monge respectivamente, entonces*

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma^*(x,y) \leq \int_{\mathcal{X}} c(x,T^*(x))d\mu(x), \tag{2.1}$$

es decir, el costo al aplicar el plan de transporte proveniente del problema de Kantorovich es menor o igual al costo asociado al mapa de transporte del problema de Monge.

Demostración. Por la proposición 2.23 tenemos que si T^* es una solución óptima al problema de Monge entonces $\gamma_{T^*} = (id \times T^*)_\sharp \mu$ es un solución factible del problema de Kantorovich.

Luego, por la definición de push-forward,

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma_T^*(x,y) &= \langle c, \gamma_{T^*}\rangle \\ &= \langle c, (id\times T^*)_\sharp\mu\rangle \\ &= \langle c\circ(id\times T^*), \mu\rangle \\ &= \int_{\mathcal{X}} c(x,T^*(x))d\mu,\end{aligned}$$

es decir, el costo de Monge obtenido con el mapa T^* es igual al costo de Kantorovich obtenido con el plan de transporte γ_{T^*}.

Finalmente, si γ^* es una solución óptima al problema de Kantorovich entonces $\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\gamma^*(x,y) = \langle c,\gamma^*\rangle \leq \langle c,\gamma_{T^*}\rangle = \int_{\mathcal{X}} c(x,T^*(x))d\mu(x)$.

□

Reparafraseando la proporción anterior, aunque el problema de Monge tenga solución T^* no hay seguridad de que $(id\times T^*)_\sharp\mu \in \Pi(\mu,\nu)$ sea una solución óptima al problema de Kantorovich, de hecho, por lo general el costo de Monge es superior al costo de Kantorovich.

El siguiente teorema probado por Yann Brenier en el año 1991 (Brenier, 1991) garantiza la existencia a los problemas de Monge y Kantorovich en un caso muy particular y prueba que la solución a ambos problemas es la misma.

Teorema 2.25 (Brenier, 1991). *En el caso donde* $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$ *y* $c(x, y) = ||x - y||^2$*, si al menos una de las dos densidades* μ *o* ν*, supongamos* μ*, tiene densidad* ρ_μ *con respecto a la medida de Lesbegue, entonces el transporte óptimo* γ *proveniente del problema de Kantorovich es único y tiene medida total sobre el gráfico* $\{(x, T(x)) \mid x \in \mathcal{X}\}$ *de un plan de transporte de Monge* $T : \mathbb{R}^n \to \mathbb{R}^n$*. Esto significa que* $\gamma = (id, T)_\sharp \mu$*.*

Además el mapa T *es el gradiente de una única función convexa* φ*, es decir,* $T = \nabla\varphi$*. Esta función* φ *está relacionada con el potencial de Kantorovich* ψ *que resuelve el problema dual 2.26 como*

$$\varphi(x) = \frac{||x||^2}{2} - \psi(x).$$

La demostración es sumamente técnica por lo que no la replicaremos en este trabajo. El lector interesado puede consultarla en el trabajo de Santambrogio (2017).

Un ejemplo de uso clásico, es que estando en las hipótesis del teorema si encontramos un mapa de transporte T que sea el gradiente de alguna función convexa entonces sabemos que T es el único transporte óptimo. Este resultado lo usamos por ejemplo cuando calculamos la distancia de Wasserstein entre dos distribuciones gaussianas univariadas (ejemplo 2.19).

2.3. Formulación dual

Este capítulo se centra en la dualización del problema de Kantorovich. La dualidad de Kantorovich desempeña un papel central en la resolución de problemas fundamentales, como los teoremas de Brenier y Gangbo-McCann sobre la existencia y unicidad de soluciones a los problemas de Monge, así como en la estabilidad de los planes y mapas de transporte óptimo. En esta sección seguiremos principalmente el libro "Optimal transport: discretization and algorithms" de Merigot y Thibert (2020) y el artículo "Optimal Transportation and Economic Applications" de Carlier (2010).

Comencemos recordando que podemos expresar la condición de que una medida positiva $\gamma \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y})$ tenga medida marginales μ y ν como

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) = \int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) d\gamma(x,y) \quad \forall \varphi \in \mathcal{C}^0(\mathcal{X}),$$
$$\int_{\mathcal{Y}} \psi(y) d\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) d\gamma(x,y) \quad \forall \psi \in \mathcal{C}^0(\mathcal{Y}).$$

A partir de esto, y utilizando la linealidad de la integral, podemos escribir las restricciones del problema de Kantorovich ($\gamma \in \Pi(\mu,\nu)$) como

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} \Big(\varphi(x) + \psi(y)\Big) d\gamma(x,y),$$

para todo par de funciones $\varphi \in \mathcal{C}^0(\mathcal{X})$ y $\psi \in \mathcal{C}^0(\mathcal{Y})$. Podemos escribir la igualdad anterior de forma compacta como $\langle \varphi \oplus \psi, \gamma \rangle = \langle \varphi \otimes 1, \mu \rangle + \langle 1 \otimes \psi, \nu \rangle$.

Si $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ es una medida de probabilidad (y por lo tanto positiva), entonces la cantidad

$$\sup_{(\varphi,\psi)\in\mathcal{C}^0(\mathcal{X})\times\mathcal{C}^0(\mathcal{Y})} \langle \varphi \otimes 1, \mu \rangle + \langle 1 \otimes \psi, \nu \rangle - \langle \varphi \oplus \psi, \gamma \rangle$$

es nula si $\gamma \in \Pi(\mu,\nu)$ y $+\infty$ en cualquier otro caso. A partir de esta observación, podemos escribir el problema de Kantorovich como un problema de optimización sin restricciones:

$$\inf_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \sup_{(\varphi,\psi)\in\mathcal{C}^0(\mathcal{X})\times\mathcal{C}^0(\mathcal{Y})} \underbrace{\langle c, \gamma \rangle}_{\text{costo de Kantorovich}} + \underbrace{\langle \varphi \otimes 1, \mu \rangle + \langle 1 \otimes \psi, \nu \rangle - \langle \varphi \oplus \psi, \gamma \rangle}_{\text{restricciones}},$$

donde c es la función de costo. Por la linealidad de la integral podemos escribir lo anterior como $\inf_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \sup_{(\varphi,\psi)\in\mathcal{C}^0(\mathcal{X})\times\mathcal{C}^0(\mathcal{Y})} \langle \varphi, \mu \rangle + \langle \psi, \nu \rangle + \langle c - \varphi \oplus \psi, \gamma \rangle$.

La dualidad de Kantorovich surge de invertir el ínfimo con el supremo:

$$\sup_{(\varphi,\psi)\in\mathcal{C}^0(\mathcal{X})\times\mathcal{C}^0(\mathcal{Y})} \inf_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \langle \varphi, \mu \rangle + \langle \psi, \nu \rangle + \langle c - \varphi \oplus \psi, \gamma \rangle.$$

Podemos simplificar la formulación anterior notando que

$$\inf_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \langle c - (\varphi \oplus \psi), \gamma \rangle = \begin{cases} 0 & \text{si } c \geq \varphi \oplus \psi, \\ -\infty & \text{si no.} \end{cases}$$

Esto se debe a que por definición

$$\langle c - \varphi \oplus \psi, \gamma \rangle = \int_{\mathcal{X}\times\mathcal{Y}} \Big[c(x,y) - (\varphi(x) + \psi(y)) \Big] d\gamma(x,y),$$

por lo que si $c \geq \varphi \oplus \psi$ entonces $c(x,y) - (\varphi(x) + \psi(y)) \geq 0$ para todo par (x,y) y por lo tanto el ínfimo será 0 tomando una medida $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ cuyo soporte no contenga los puntos donde $c(x,y) - (\varphi(x) - \psi(y)) > 0$. Por otro lado, si $c < \varphi \oplus \psi$ entonces $c(x,y) - (\varphi(x) + \psi(y)) < 0$ y el ínfimo será $+\infty$ tomando una medida adecuada.

Juntando todo lo anterior, el problema dual de Kantorovich es

Definición 2.26 (Problema Dual de Kantorovich). *Dados dos espacios métricos compactos $\mathcal{X}$ e $\mathcal{Y}$, dos medidas de probabilidad $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$ y una función de costo $c \in \mathcal{C}^0(\mathcal{X} \times \mathcal{Y})$, el problema Dual de Kantorovich es el problema de optimización:*

$$(\varphi^*, \psi^*) \quad = \quad \underset{\substack{\varphi \in \mathcal{C}^0(\mathcal{X}), \psi \in \mathcal{C}^0(\mathcal{Y}) \\ \varphi \oplus \psi \leq c}}{\arg\sup} \Big\{ \int_{\mathcal{X}} \varphi(x) d\mu(x) - \int_{\mathcal{Y}} \psi(y) d\nu(y) \Big\}. \qquad (DP)$$

Cuando tenemos un problema primal y uno dual es natural que nos preguntemos que relación, si es que existe, tienen ambos problemas de optimización. La siguiente proposición da una respuesta parcial a dicha interrogante.

Proposición 2.27. *El costo asociado a una solución óptima del problema de Kantorovich primal es mayor o igual al costo asociado a una solución óptima del problema dual. Abusando de la notación podemos escribir $(KP) \geq (DP)$.*

Demostración. Sean $\varphi \in \mathcal{C}^0(\mathcal{X})$, $\psi \in \mathcal{C}^0(\mathcal{Y})$ y $\gamma \in \Pi(\mu, \nu)$ satisfaciendo la restricción $\varphi \oplus \psi \leq c$. Como $\gamma \in \Pi(\mu, \nu)$ tenemos que

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) = \int_{\mathcal{X}\times\mathcal{Y}} \varphi(x) d\gamma(x,y),$$

y

$$\int_{\mathcal{Y}} \psi(y) d\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) d\gamma(x,y).$$

Luego

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) - \int_{\mathcal{Y}} \psi(y) d\nu(y) = \int_{\mathcal{X}\times\mathcal{Y}} \Big(\varphi(x) - \psi(y) \Big) d\gamma(x,y),$$

equivalentemente de forma compacta, $\langle \varphi, \mu \rangle - \langle \psi, \nu \rangle = \langle \varphi \ominus \psi, \gamma \rangle$.

Finalmente, por la restricción $\varphi(x) - \psi(y) \leq c(x, y)$ tenemos que

$$\langle \varphi, \mu \rangle - \langle \psi, \nu \rangle = \langle \varphi \ominus \psi, \gamma \rangle \leq \langle c, \gamma \rangle,$$

para cualquier combinación de φ, ψ y γ tales que se cumpla la restricción $\varphi(x) - \psi(y) \leq c(x, y)$, por lo tanto

$$(DP) = \sup_{\varphi \oplus \psi \leq c} \langle \varphi, \mu \rangle - \langle \psi, \nu \rangle \leq \inf_{\gamma \in \Pi(\mu, \nu)} \langle c, \gamma \rangle = (KP).$$

□

Para probar la existencia de las soluciones al problema dual debemos introducir el concepto de c-transformación. Comencemos con una idea intuitiva de este concepto: el problema de Kantorovich dual consiste en hallar $\varphi \in \mathcal{C}^0(\mathcal{X})$ y $\psi \in \mathcal{C}^0(\mathcal{Y})$ tales que maximicen la cantidad

$$\int_{\mathcal{X}} \varphi(x) d\mu(x) - \int_{\mathcal{Y}} \psi(y) d\nu(y),$$

bajo la restricción de que $\varphi \oplus \psi \leq c$. Un camino para encontrar tales φ y ψ es por ejemplo fijar ψ (y por lo tanto $\langle \psi, \nu \rangle$ quedará fijo) y encontrar φ que maximice $\int_{\mathcal{X}} \varphi d\mu$ pero cumpliendo además la restricción $\varphi \oplus \psi \leq c$. Con ψ fija, podemos escribir esta restricción como

$$\varphi(x) \leq \inf_{y \in \mathcal{Y}} c(x, y) - \psi(y) \quad \forall x \in \mathcal{X}.$$

Como queremos elegir la función φ más grande posible, debemos quedarnos con aquella donde se da la igualdad, es decir, $\varphi(x) = \min_{y \in \mathcal{Y}} c(x, y) - \psi(y) \quad \forall x \in \mathcal{X}$. Esto nos permite replantear el problema dual de Kantorovich como

$$\begin{aligned} &\underset{\substack{\varphi \in \mathcal{C}^0(\mathcal{X}), \psi \in \mathcal{C}^0(\mathcal{Y}) \\ \varphi \oplus \psi \leq c}}{\arg\sup} \Big\{ \int_{\mathcal{X}} \varphi d\mu - \int_{\mathcal{Y}} \psi d\mu \Big\} = \\ &= \underset{\psi \in \mathcal{C}^0(\mathcal{Y})}{\arg\sup} \Big\{ \int_{\mathcal{X}} \Big(\inf_{y \in \mathcal{Y}} c(x, y) - \psi(y) \Big) d\mu(x) - \int_{\mathcal{Y}} \psi(y) d\nu(y) \Big\}. \end{aligned}$$

Podríamos haber comenzado al revés, es decir, fijar φ y tendríamos que $\psi(y) = \sup_{x \in \mathcal{X}} -c(x, y) + \varphi(x) \ \forall y \in \mathcal{Y}$. A continuación introducimos la no-

tación de c-transformación, la cual es una manera compacta de escribir lo anterior.

Definición 2.28 (c-transformación). *La c-transformación (resp. $\bar{c}$-transformación) de una función $\psi : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ (resp. $\varphi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$) se define como*

$$\psi^c : x \in \mathcal{X} \mapsto \inf_{y \in Y} c(x, y) + \psi(y)$$
$$\varphi^{\bar{c}} : y \in \mathcal{Y} \mapsto \sup_{x \in \mathcal{X}} -c(x, y) + \varphi(x).$$

A partir de la c-transformación podemos reformular el problema dual de Kantorovich como un problema de maximización sin restricciones

$$\arg\sup_{\psi \in \mathcal{C}^0(\mathcal{Y})} \int_{\mathcal{X}} \psi^c d\mu - \int_{\mathcal{Y}} \psi d\nu.$$

Lo anterior motiva la siguiente definición:

Definición 2.29 (Funcional de Kantorovich). *El funcional de Kantorovich definido en $\mathcal{C}^0(\mathcal{Y})$ es*

$$\mathcal{K}(\psi) = \int_{\mathcal{X}} \psi^c d\mu - \int_{\mathcal{Y}} \psi d\nu.$$

Con esta notación, el problema dual de Kantorovich se puede escribir

$$\arg\sup_{\psi \in \mathcal{C}^0(\mathcal{Y})} \mathcal{K}(\psi).$$

Estamos ahora en condiciones de enunciar la existencia de las soluciones al problema dual.

Teorema 2.30 (Existencia de potenciales duales). *El problema dual (DP) siempre admite una solución, la cual se puede asumir de la forma (φ, ψ) con $\varphi = \psi^c$ y $\psi = \varphi^{\bar{c}}$.*

Más aún, tenemos el siguiente resultado que nos permite trabajar con el problema dual para resolver el problema primal.

Teorema 2.31 (Dualidad fuerte). *Sean $\mathcal{X}$ e $\mathcal{Y}$ dos espacios métricos compactos y $c \in \mathcal{C}^0(\mathcal{X} \times \mathcal{Y})$ una función de costo. El máximo se alcanza en el problema dual (DP) y además dicho máximo es igual al mínimo costo del problema primal (KP).*

Las demostraciones de estos teoremas exceden de los objetivos de esta tesis, pero los lectores interesados pueden consultarlas en Merigot y Thibert (2020). En un trabajo más reciente, Staudt et al. (2022), prueban la unicidad de los potenciales de Kantorovich con hipótesis más relajadas que las presentadas en esta tesis.

Observación 2.32. *Las condiciones de optimilidad de los problemas primal y dual permiten probar que si $\gamma \in \Pi(\mu, \nu)$ es una solución óptima del problema de Kantorovich primal entonces*

$$sop(\gamma) \subset \Big\{(x, y) \in \mathcal{X} \times \mathcal{Y} \;\; | \;\; \varphi(x) + \psi(y) = c(x, y)\Big\}.$$

2.4. Regularización

El problema de transporte óptimo, tal como fue formulado por Kantorovich, ha sido un pilar en el análisis y aplicación de la teoría de transporte en diversas áreas. Sin embargo, a pesar de su amplia utilidad teórica, el problema clásico enfrenta desafíos significativos al utilizarlo en aplicaciones con datos reales, especialmente en términos de complejidad computacional y sensibilidad a perturbaciones de los datos. Es curioso que la regularización en el contexto del transporte óptimo se propuso hace relativamente poco tiempo, uno de los primeros trabajos fue el de Cuturi (2013). A continuación presentamos un ejemplo para entender la sensibilidad en la peturbación de los datos.

Ejemplo 2.33. *Sea $\epsilon > 0$ tan pequeño como queramos. Supongamos que $\mathcal{X}_1 = \{(0.5 - \epsilon, 0), (0.5, 2)\}$ e $\mathcal{Y} = \{(0, 1), (1, 1)\}$ son dos conjuntos de puntos en el plano y que tenemos dos medidas $\mu = \frac{2}{3}\delta_{(0.5-\epsilon,0)} + \frac{1}{3}\delta_{(0.5,2)}$ y $\nu = \frac{2}{3}\delta_{(0,1)} + \frac{1}{3}\delta_{(1,1)}$. Usaremos como función de costo la distancia euclídea. En este caso el plan de transporte óptimo de Kantorovich es un transporte óptimo de Monge, específicamente*

$$T_1((0.5 - \epsilon, 0)) = (0, 1) \quad y \quad T_1((0.5, 2)) = (1, 1).$$

Por otro lado, si al momento de adquirir los datos hubiésemos tenido que $\mathcal{X}_2 = \{(0.5 + \epsilon, 0), (0.5, 2)\}$ entonces la medida μ cambia a $\mu = \frac{2}{3}\delta_{(0.5+\epsilon,0)} + \frac{1}{3}\delta_{(0.5,2)}$. En este caso, el plan de transporte óptimo de Kantorovich no es un transporte de Monge ya que necesitamos dividir la masa ubicada en $(0.5 + \epsilon, 0)$

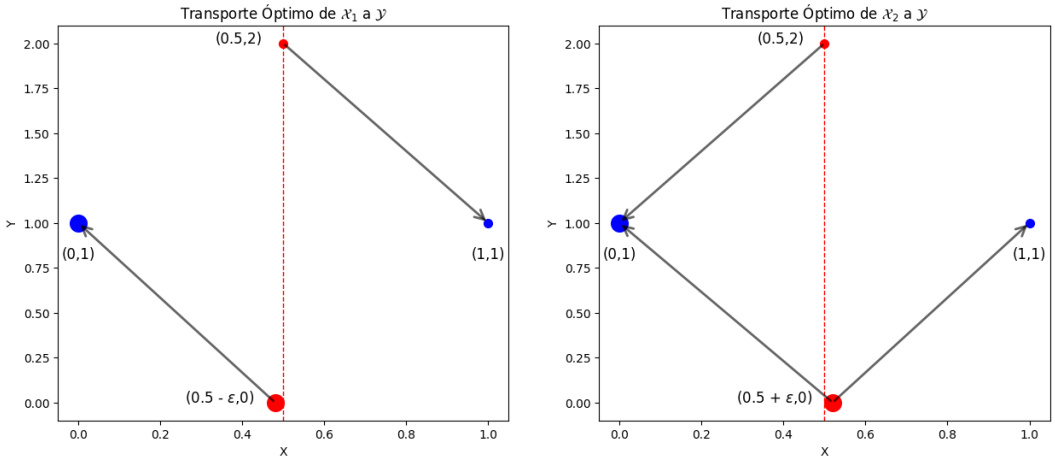


Figura 2.21: Representación de la situación del ejemplo 2.33. De izquierda a derecha: transporte óptimo entre $\mathcal{X}_1$ e $\mathcal{Y}$ y transporte óptimo entre $\mathcal{X}_2$ e $\mathcal{Y}$ respectivamente. Una pequeña perturbación en un dato puede cambiar radicalmente el transporte óptimo.

y transportar $\frac{1}{3}$ *de la misma a* $(0,1)$ *y el otro* $\frac{1}{3}$ *a* $(1,1)$. *La figura 2.21 muestra ambas situaciones.*

La sensibilidad en los datos es un problema al momento de querer aplicar los conceptos del transporte óptimo en la práctica porque dificulta la reproducibilidad de los resultados. Por otro lado, la complejidad computacional también puede llegar a ser un impedimento en la práctica, cuando se tiene muchos datos. Estos desafíos motivan la consideración de una versión regularizada del problema, que no solo mejora la tratabilidad computacional sino también la estabilidad de las soluciones obtenidas. Por todas estas razones, la exploración del problema de Kantorovich regularizado no solo es una extensión lógica del transporte óptimo tradicional sino una necesidad para avanzar en la aplicación de estas teorías a problemas modernos de gran escala y alta complejidad.

El problema de Kantorovich regularizado introduce un término de regularización en la función objetivo clásica, que típicamente penaliza la entropía de los planes de transporte. Este enfoque tiene múltiples ventajas, por ejemplo desde una perspectiva computacional, la regularización entrópica transforma el problema en uno más suave, facilitando el uso de algoritmos eficientes como el algoritmo Sinkhorn-Knopp (capítulo 3), el cual es un algoritmo iterativo y donde se puede aprovechar técnicas de optimización modernas para acelerar significativamente la convergencia y así obtener un plan de transporte apro-

ximado más rápido que uno exacto. Esto se vuelve imprescindible al trabajar con grandes volúmenes de datos. Al ajustar el parámetro de regularización, podemos controlar el grado de regularidad en la solución, permitiendo un balance entre la precisión del modelo original y la robustez frente a variaciones en los datos.

Presentamos a continuación el problema de transporte óptimo regularizado:

Definición 2.34 (Regularización). *Sean $\mathcal{X}$ e $\mathcal{Y}$ dos espacios métricos compactos, $\mu \in \mathcal{P}(\mathcal{X})$ y $\nu \in \mathcal{P}(\mathcal{Y})$ dos medidas de probabilidad, $c \in \mathcal{C}^0(\mathcal{X} \times \mathcal{Y})$ una función de costo no negativa y $\epsilon > 0$. El problema de transporte óptimo de Kantorovich regularizado (KRP_ϵ) es el siguiente problema de optimización*

$$\gamma_\epsilon^* = \underset{\gamma \in \Pi(\mu,\nu)}{\arg\inf} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y) + \epsilon D_{KL}(\gamma || \mu \otimes \nu) \right\} \qquad (KRP_\epsilon)$$

Observación 2.35. *Al calcular la divergencia de Kullback-Leiber $D_{KL}(\gamma||\mu \otimes \nu)$ utilizaremos la convención $d\mu \otimes d\nu(x,y) := d\mu(x)d\nu(y)$.*

La idea base de por qué utilizar una regularización entrópica es que la misma fuerza a la solución a tener un soporte disperso en contraste con que los planes de transportes provenientes del problema de Kantorovich tienen soporte concentrado en una gráfica $\{(x, T(x))\}_{x\in\mathbb{R}}$ como prueba el teorema de Brenier (teorema 2.25). Una consecuencia de esto es que ayuda a estabilizar el cálculo ya que el problema de Kantorovich regularizado con entropía define un problema fuertemente convexo (a diferencia del problema de Kantorovich que es convexo). Una segunda consecuencia, más importante que la anterior, es que la solución al problema regularizado por entropía es un 'escalamiento diagonal' de e^{-C} donde C es la matriz de costo y además existe un algoritmo eficiente para encontrar dicha solución: el algoritmo de Sinkhorn-Knopp, el cual presentaremos en la sección 3.2.

Es importante notar que cuando $\epsilon \to 0$ recuperamos el problema de Kantorovich mientras que al aumentar lo suficiente ϵ, el termino dominante es $\epsilon D_{KL}(\gamma||\mu \otimes \nu)$ y el plan de transporte resultante es aproximadamente $\mu \otimes \nu$, es decir, $\gamma^* \approx \mu \otimes \nu$. El problema de Kantorovich regularizado abre la puerta a nuevas interpretaciones y aplicaciones en campos como la mecánica estadística (Taskesen et al. 2022), el machine learning (Paty y Cuturi, 2020) y la economía (Galichon, 2016), donde conceptos como la entropía y la regularización juegan roles fundamentales.

Capítulo 3

Implementación del transporte óptimo

Después de haber explorado en detalle los fundamentos teóricos del transporte óptimo, así como la formulación dual del mismo y la técnica de regularización, este capítulo se centrará en la implementación práctica de estos conceptos. A través de una combinación de algoritmos, técnicas de discretización y aplicaciones reales, proporcionaremos una guía para llevar los conceptos teóricos a la práctica.

El primer paso en la implementación del transporte óptimo consiste en la discretización del problema. La discretización permite convertir problemas continuos en problemas manejables computacionalmente, facilitando así su resolución. En la primera sección cubriremos diferentes enfoques y técnicas de discretización.

Una vez discretizado el problema, nos proponemos introducir dos de los algoritmos mas utilizados para resolver el problema de Monge y el transporte óptimo regularizado: la asignación lineal y el algoritmo de Sinkhorn-Knopp respectivamente.

Finalmente explicaremos cómo aplicar las nociones de transporte óptimo a problemas de aprendizaje automático que involucran imágenes. Mostraremos el estado actual del transporte óptimo dentro del aprendizaje automático, donde el mismo está ganando cada vez más relevancia.

3.1. Discretización

En esta sección, examinaremos cómo discretizar tanto el problema de Kantorovich como el problema regularizado. Esta discretización es fundamental para desarrollar algoritmos que permitan implementar estas ideas en una computadora para mayor eficiencia. Seguiremos principalmente el trabajo de Merigot y Thibert (2020).

Problema de Kantorovich

Veamos primero la discretización del problema de Kantorovich primal (KP). Supondremos que $\mathcal{X}$ e $\mathcal{Y}$ son espacios métricos finitos, μ y ν son distribuciones discretas,

$$\mu = \sum_{x\in\mathcal{X}} \mu_x \delta_x, \quad \nu = \sum_{y\in\mathcal{Y}} \nu_y \delta_y,$$

donde cada μ_x y ν_y son no negativos y $\sum_{x\in\mathcal{X}} \mu_x = \sum_{y\in\mathcal{Y}} \nu_y = 1$. En este contexto, el conjunto de planes de transportes es

$$\Pi(\mu,\nu) = \Big\{\gamma = \sum_{x,y} \gamma_{xy}\delta_{(x,y)} \quad | \quad \gamma_{xy} \geq 0, \sum_{y\in\mathcal{Y}} \gamma_{xy} = \mu_x, \sum_{x\in\mathcal{X}} \gamma_{x,y} = \nu_y\Big\}.$$

Como los espacios $\mathcal{X}$ e $\mathcal{Y}$ son finitos, podemos representar un plan de transporte γ como una matriz

$$P_\gamma = (\gamma_{xy})_{xy},$$

donde podemos interpretar que el elemento γ_{xy} representa la cantidad de masa transportada desde la posición x a la posición y. Desde un punto de vista probabilístico, la matriz P_γ es una distribución conjunta discreta.

Para ilustrar esta interpretación revisitemos el ejemplo 2.33. Si consideramos los espacios $\mathcal{X}_1 = \{(0.5-\epsilon, 0), (0.5, 2)\}$ e $\mathcal{Y} = \{(0,1),(1,1)\}$ y las medidas $\mu = \frac{2}{3}\delta_{(0.5-\epsilon,0)} + \frac{1}{3}\delta_{(0.5,2)}$ y $\nu = \frac{2}{3}\delta_{(0,1)} + \frac{1}{3}\delta_{(1,1)}$ entonces el plan de transporte óptimo es

$$\gamma_1^* = \begin{pmatrix} 2/3 & 0 \\ 0 & 1/3 \end{pmatrix}.$$

Esto quiere decir que de la masa total de μ, la $2/3$ parte se debe transportar desde el punto $(0.5-\epsilon, 0)$ hasta el punto $(0,1)$, mientras que el restante $1/3$

debe transportarse desde la posición $(0.5, 2)$ hasta $(1, 1)$.

Por otro lado, si consideramos los conjuntos $\mathcal{X}_2 = \{(0.5 + \epsilon, 0), (0.5, 2)\}$ e $\mathcal{Y} = \{(0, 1), (1, 1)\}$, entonces el plan de transporte óptimo es

$$\gamma_2^* = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & 0 \end{pmatrix},$$

lo cual podemos interpretar como que para "llenar" la masa total del punto $(0, 1)$ debemos transportar $\frac{1}{3}$ desde la posición $(0.5, 2)$ y otro $\frac{1}{3}$ de la posición $(0.5 + \epsilon, 0)$, mientras que la masa del punto $(1, 1)$ se llena con el restante $\frac{1}{3}$ de la posición $(0.5 + \epsilon, 0)$.

Además, por ser $\mathcal{X}$ e $\mathcal{Y}$ espacios discretos tenemos que la función de costo $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ también se puede representar con una matriz:

$$C = (c(x, y))_{xy}.$$

A partir de estas definiciones podemos definir el problema de Kantorivch discreto.

Definición 3.1 (Problema de Kantorovich discreto). *El problema de Kantorovich discreto es el problema de optimización*

$$P_\gamma^* = \arg\min_{\gamma} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} c(x, y)\gamma_{xy} = \min_{P_\gamma} \langle C, P_\gamma \rangle_F \qquad (KPD)$$

restricto a $\gamma \in \Pi(\mu, \nu)$, donde $\langle \cdot, \cdot \rangle_F$ es el producto interno de Frobenius.

Recordamos que el problema dual consiste en encontrar una función $\psi : \mathcal{Y} \to \mathbb{R}$ que maximice el funcional de Kantorovich (ver 2.29). En el caso discreto, este se convierte en

$$\mathcal{K}(\psi) = \sum_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} \Big(c(x, y) + \psi(y) \Big) \mu_x - \sum_{y \in \mathcal{Y}} \psi(y) \nu_y.$$

A partir de esto, el problema de Kantorovich dual discreto es

Definición 3.2 (Problema de Kantorovich dual discreto). *El problema de Kantorovich dual discreto (DPD) es el siguiente problema de optimización*

$$\psi^* \quad = \quad \arg\max_{\psi} \mathcal{K}(\psi) \qquad (DPD),$$

donde $\psi : \mathcal{Y} \to \mathbb{R}$.

Así como en el caso continuo tenemos el teorema de dualidad fuerte (teorema 2.31), en el caso discreto tenemos un equivalente y por lo tanto

$$\min_{\gamma \in \Pi(\mu,\nu)} \langle C, P_\gamma \rangle_F = \max_{\psi : \mathcal{Y} \to \mathbb{R}} \mathcal{K}(\psi).$$

Problema de Kantorovich regularizado

Así como en la versión continua utilizamos la divergencia de Kullback-Leibler como regularizador, en el caso discreto utilizaremos la entropía, la cual introduciremos a continuación.

Definición 3.3 (Entropía de un plan de transporte). *Definimos la entropía de un plan de transporte* $\gamma \in \Pi(\mu, \nu)$ *como* $H(\gamma) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} h(\gamma_{x,y})$ *donde*

$$h(t) = \begin{cases} tlog(t) & si\ t > 0, \\ 0 & si\ t = 0. \end{cases}$$

Observamos que esto coincide con la definición de entropía dada en el capítulo 1 teniendo en cuenta que γ es una distribución conjunta. A partir de la entropía definimos el problema de Kantorovich regularizado como

Definición 3.4. *Sean* $\mathcal{X}$ *e* $\mathcal{Y}$ *dos espacios métricos compactos discretos,* μ *y* ν *dos medidas discretas sobre* $\mathcal{X}$ *e* $\mathcal{Y}$ *respectivamente,* C *una matriz de costo entre los puntos de* $\mathcal{X}$ *e* $\mathcal{Y}$ *y* $\epsilon > 0$ *constante. El problema de Kantorovich regularizado discreto es el problema de optimización*

$$P^*_{\gamma,\epsilon} \quad = \quad \arg\min_{P_\gamma \in \Pi(\mu,\nu)} \langle C, P_\gamma \rangle_F + \epsilon H(\gamma) \qquad (KPD_\epsilon).$$

El siguiente teorema nos garantiza la existencia y unicidad de la solución del problema regularizado. Su demostración es sumamente técnica por lo que la omitiremos y se puede consultar por ejemplo en el trabajo de Merigot y Thibert (2020).

Teorema 3.5. *Sean* $\mathcal{X}$ *e* $\mathcal{Y}$ *dos espacios métricos finitos y* $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ *dos medidas de probabilidad en* $\mathcal{X}$ *e* $\mathcal{Y}$ *respectivamente. El problema*

regularizado de Kantorovich (KPD_ϵ) *tiene una única solución* $P^*_{\gamma,\epsilon} \in \Pi(\mu,\nu)$. *Más aún, si* μ_x *y* ν_y *son todos positivos entonces* $\gamma_{x,y} > 0 \ \ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}$.

Al igual que el problema de Kantorovich, el problema regularizado también tiene una contraparte dual. Comencemos definiendo el Langragiano de (KPD_ϵ):

$$L(\gamma,\varphi,\psi) := \langle C, P_\gamma \rangle_F + \epsilon H(\gamma) + \sum_{x\in\mathcal{X}} \varphi(x)\Big(\mu_x - \sum_{y\in\mathcal{Y}} \gamma_{x,y}\Big) + \sum_{y\in\mathcal{Y}} \psi(y)\Big(\sum_{x\in\mathcal{X}} \gamma_{x,y} - \nu_y\Big),$$

donde $\varphi : \mathcal{X} \to \mathbb{R}$ y $\psi : \mathcal{Y} \to \mathbb{R}$ son multiplicadores de Langrange. Utilizando este Langragiano, el problema regularizado puede escribirse como

$$(KPD_\epsilon) = \min_{\gamma\in\Pi(\mu,\nu)} \sup_{\varphi,\psi} L(\gamma,\varphi,\psi).$$

Como antes, el problema dual se obtiene intercambiando el mínimo y el supremo, si además simplificamos obtenemos

$$\sup_{\varphi,\psi} \min_{\gamma\in\Pi(\mu,\nu)} L(\gamma,\varphi,\psi) = \sup_{\varphi,\psi} \min_{\gamma\in\Pi(\mu,\nu)} \sum_{x,y} \gamma_{x,y}\Big[c(x,y) + \psi(y) - \varphi(x) + \epsilon\Big(log(\gamma_{x,y})\Big)\Big] + \sum_x \varphi(x)\mu_x - \sum_y \psi(y)\nu_y.$$

Si derivamos $L(\gamma,\varphi,\psi)$ respecto a $\gamma_{x,y}$ obtenemos que para φ y ψ fijos el plan de transporte óptimo debe satisfacer

$$c(x,y) + \psi(y) - \varphi(x) + \epsilon \log(\gamma_{x,y}) = 0,$$

de donde

$$\gamma_{x,y} = e^{\frac{1}{\epsilon}\Big(\varphi(x) - \psi(y) - c(x,y)\Big)}.$$

Juntando todo lo anterior tenemos lo siguiente:

Definición 3.6 (Problema regularizado dual). *El dual del problema de Kantorovich regularizado está definido por*

$$(\varphi^*_\epsilon, \psi^*_\epsilon) \quad = \quad \arg\sup_{\varphi,\psi} \mathcal{K}_\epsilon(\varphi,\psi), \qquad (DPD_\epsilon)$$

siendo

$$\mathcal{K}_\epsilon(\varphi,\psi) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \epsilon e^{\frac{1}{\epsilon}\left(\varphi(x)-\psi(y)-c(x,y)\right)} + \sum_{x\in\mathcal{X}}\varphi(x)\mu_x - \sum_{y\in\mathcal{Y}}\psi(y)\nu_y.$$

Nuevamente tenemos un resultado que relaciona las soluciones del problema de Kantorovich regularizado con su dual.

Teorema 3.7 (Dualidad fuerte). *El máximo del problema dual se alcanza, es decir, existen $\varphi : \mathcal{X} \to \mathbb{R}$ y $\psi : \mathcal{Y} \to \mathbb{R}$ tales que*

$$(KPD_\epsilon) = (DPD_\epsilon) = \mathcal{K}^\epsilon(\varphi,\psi).$$

Como corolario tenemos:

Corolario 3.8. *Si φ y ψ son la solución al problema dual (DPD_ϵ), entonces la solución al problema primal (KPD_ϵ) está dado por*

$$\gamma_{x,y} = e^{\frac{1}{\epsilon}\left(\varphi(x)-\psi(y)-c(x,y)\right)} = e^{\varphi(x)/\epsilon}e^{-c(x,y)/\epsilon}e^{-\psi(x)/\epsilon}.$$

Una manera natural de maximizar $\mathcal{K}_\epsilon(\varphi,\psi)$ es maximizar alternadamente φ y ψ, donde tenemos garantizada la convergencia al máximo global dado que $\mathcal{K}_\epsilon(\varphi,\psi)$ es una función cóncava y C^1. Con esto en mente, Cuturi (2013) propone utilizar el algoritmo de Sinkhorn-Knopp, que es un algoritmo iterativo, para encontrar la solución al problema regularizado por entropía. En el caso donde utilizamos una regularización por entropía, cada problema de maximización parcial (en φ y ψ) tiene una solución explícita, las cuales están relacionadas con la noción de c-transformación:

Proposición 3.9. *1. Dada $\psi : \mathcal{Y} \to \mathbb{R}$, el maximizador de $\mathcal{K}_\epsilon(\cdot,\psi)$ se alcanza en una única función $\varphi : \mathcal{X} \to \mathbb{R}$, denotada ψ_ϵ^c y definida por*

$$\psi_\epsilon^c(x) = \epsilon\log(\mu_x) - \epsilon\log\Big(\sum_{y\in\mathcal{Y}} e^{\frac{1}{\epsilon}(-c(x,y)-\psi(y))}\Big).$$

2. Dada $\varphi : \mathcal{X} \to \mathbb{R}$, el maximizador de $\mathcal{K}_\epsilon(\varphi,\cdot)$ se alcanza en una única función $\psi : \mathcal{Y} \to \mathbb{R}$, denotada $\varphi_\epsilon^{\bar{c}}$ y definida por

$$\varphi_\epsilon^{\bar{c}}(y) = -\epsilon\log(\nu_y) + \epsilon\log\Big(\sum_{x\in\mathcal{X}} e^{\frac{1}{\epsilon}(-c(x,y)+\varphi(x))}\Big).$$

Demostración. Derivando $\mathcal{K}_\epsilon(\cdot, \psi)$ con respecto a $\varphi(x)$ obtenemos

$$\mu_x = e^{\frac{\varphi(x)}{\epsilon}} \sum_{y \in \mathcal{Y}} e^{\frac{1}{\epsilon}(-c(x,y)-\psi(y))},$$

lo cual implica la igualdad de (1). El punto (2) se prueba de forma análoga. □

A las funciones definidas en el teorema anterior las llamaremos c-transformación regularizada:

Definición 3.10 (c-transformación regularizada). *Dado $\psi : \mathcal{Y} \to \mathcal{X}$, diremos que la función ψ_ϵ^c es la c-transformación regularizada. Por otro lado, dado $\varphi : \mathcal{X} \to \mathbb{R}$, llamaremos a $\varphi_\epsilon^{\overline{c}}$ como $\overline{c}$-transformación regularizada.*

La razón de la elección de este nombre queda claro una vez que notamos que

$$\begin{aligned}
\lim_{\epsilon \to 0} \psi_\epsilon^c(x) &= \lim_{\epsilon \to 0} \epsilon \Big[\log(\mu_x) - \log\Big(\sum_{y \in \mathcal{Y}} e^{\frac{1}{\epsilon}(-c(x,y)-\psi(y))} \Big) \Big] \\
&= \min_{y \in \mathcal{Y}} c(x,y) + \psi(y) \\
&= \psi^c(x),
\end{aligned}$$

es decir, cuando $\epsilon \to 0$, recuperamos la $c-$transformación como la vimos en el capítulo anterior.

3.2. Algoritmos

En esta sección presentamos los algoritmos utilizados en la práctica para calcular el transporte óptimo, tanto en la versión de Kantorovich como en la versión regularizada. Nos enfocaremos en dos algoritmos claves: el algoritmo de asignación lineal y el algoritmo de Sinkhorn-Knopp. El algoritmo de asignación lineal, también conocido como algoritmo húngaro, resuelve el problema de asignación minimizando el costo total de transporte entre fuentes y destinos. Por otro lado, el algoritmo de Sinkhorn-Knopp, una técnica iterativa eficiente, se utiliza para calcular soluciones regularizadas.

3.2.1. Algoritmo Húngaro

En el caso particular donde los conjuntos $\mathcal{X}$ e $\mathcal{Y}$ tienen mismo cardinal finito N y las medidas μ y ν son distribuciones uniformes

$$\mu = \frac{1}{N}\sum_{x\in\mathcal{X}} \delta_x, \quad \nu = \frac{1}{N}\sum_{y\in\mathcal{Y}} \delta_y,$$

el problema de Monge se corresponde con el problema lineal llamado problema de asignación lineal, el cuál es sumamente conocido en optimización combinatoria:

$$\text{mín}\left\{\frac{1}{N}\sum_{x\in\mathcal{X}} c(x,\sigma(x)) \;\;|\;\; \sigma : \mathcal{X}\to\mathcal{Y} \text{ es una biyección}\right\}. \qquad (AP)$$

Observación 3.11. *Si el cardinal de $\mathcal{X}$ e $\mathcal{Y}$ es N entonces la cantidad de biyecciones $\sigma : \mathcal{X}\to\mathcal{Y}$ es $N!$, lo cual crece muy rápido cuando la cantidad de puntos. Esto hace que no sea práctico encontrar la solución probando las biyecciones una por una.*

Nos proponemos mostrar que el problema de asignación coincide en este caso con el problema de Kantorovich, para lo cual necesitamos algunos resultados previos.

A continuación probaremos que las soluciones al problema de Kantorovich y el problema de Asignación coinciden cuando $\mathcal{X}$ e $\mathcal{Y}$ tienen el mismo cardinal y además las medidas μ y ν son uniformes.

Teorema 3.12. *Sean μ y ν medidas de probabilidad uniformes en $\mathcal{X}$ e $\mathcal{Y}$ espacios métricos finitos de mismo cardinal, entonces $(AP) = (KP)$.*

Demostración. Sean $\{x_1,\dots,x_N\}$ y $\{y_1,\dots,y_N\}$ ordenes cualquiera de $\mathcal{X}$ e $\mathcal{Y}$. Luego, γ es un plan de transporte entre μ y ν si y sólo si $P_\gamma \in \mathcal{B}_N$. Como $\mathcal{G}_N \subset \mathcal{B}_N$ entonces $(KP) \leq (AP)$. Por otro lado, el mínimo del problema de Kantorovich (KP) se alcanza en un vértice del poliedro $\mathcal{B}_N$, que por el teorema de Birkhoff (teorema 1.35), es una matriz de permutación, por lo tanto $(KP) \geq (AP)$, obteniéndose de esta manera la igualdad. □

Un algoritmo clásico para resolver el problema de la asignación lineal es el conocido "Algoritmo Húngaro" propuesto en el artículo "The Hungarian method for the assignment problem"de Harold Kuhn (1955). Dos años después,

James Munkres refino el algoritmo y esta versión es la que suele usarse en la actualidad (Munkres, 1957). A continuación presentamos el pseudo código del algoritmo Húngaro en la versión de Munkres.

Algorithm 1 Algoritmo Húngaro para el problema de asignación lineal

Require: Matriz de costos $C = [c_{ij}]$ de dimensión $n \times n$
Ensure: Asignación óptima de mínimo costo
Paso 1: Reducción de filas
for cada fila i en C **do**
Restar el valor mínimo de la fila i de todos los elementos en esa fila
end for
Paso 2: Reducción de columnas
for cada columna j en C **do**
Restar el valor mínimo de la columna j de todos los elementos en esa columna
end for
Paso 3: Cubrir ceros con el mínimo número de líneas
Cubrir todos los ceros en la matriz resultante utilizando el menor número de líneas horizontales y verticales
if el número de líneas es igual a n **then**
Ir al paso 5
else
Ir al paso 4
end if
Paso 4: Ajuste de la matriz
Encontrar el valor mínimo no cubierto por ninguna línea
Restar este valor de todos los elementos no cubiertos y sumarlo a los elementos cubiertos por dos líneas
Ir al paso 3
Paso 5: Encontrar asignación óptima
Utilizar los ceros en la matriz resultante para encontrar una asignación óptima

Presentamos a continuación un ejemplo simple para entender el funcionamiento del algoritmo 1.

Ejemplo 3.13. *Supongamos que* $\mathcal{X} = \{(0,0),(2,0)\}$ *e* $\mathcal{Y} = \{(1,1),(4,1)\}$. *En el problema de asignación lineal necesitamos que las medidas sean uniforme sobre los puntos, por lo tanto*

$$\mu = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(2,0)} \quad y \quad \nu = \frac{1}{2}\delta_{(1,1)} + \frac{1}{2}\delta_{(4,1)}.$$

Utilizando como función de costo el cuadrado de la distancia euclídea te-

nemos que la matriz de costo C es

$$C = \begin{pmatrix} 2 & 17 \\ 2 & 5 \end{pmatrix}.$$

Paso 1: Reducción de filas *El mínimo de ambas filas es* 2*. Luego de restar este valor a cada fila obtenemos la matriz:*

$$\begin{pmatrix} 0 & 15 \\ 0 & 3 \end{pmatrix}.$$

Paso 2: Reducción de columnas *El mínimo de la primera columna es* 0 *mientras que el de la segunda columna* 3*. Restando estos valores a la primera y segunda columna respectivamente obtenemos la matriz:*

$$\begin{pmatrix} 0 & 12 \\ 0 & 0 \end{pmatrix}.$$

Paso 3: *Cubrimos los ceros*

$$\begin{pmatrix} | & 12 \\ | & - \end{pmatrix}.$$

*El número de líneas es igual al número de filas (*2*), por lo que podemos proceder a la asignación óptima.*

Paso 5: *Seleccionamos los ceros independientes:*

- *Asignamos el primer punto source (0,0) al primer punto target (1,1) con costo 2.*
- *Asignamos el segundo punto source (2,0) al segundo punto target (4,1) con costo 5.*

La figura 3.1 muestra la disposición de los puntos así como la asignación óptima.

3.2.2. Algoritmo de Sinkhorn-Knopp

Anteriormente vimos en el colorario 3.8 que la solución al problema dual de Kantorovich con regularización (KPD_ϵ) tiene una forma específica. Presentamos a continuación un algoritmo ampliamente utilizado para encontrar

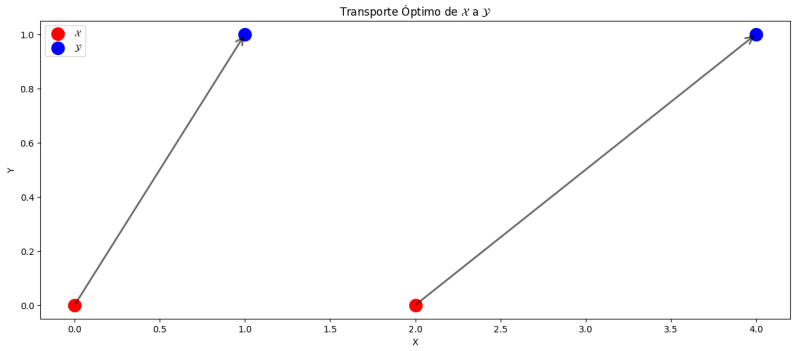


Figura 3.1: Situación descrita en el ejemplo 3.13.

dicha solución, denominado *algoritmo de Sinkhorn-Knopp*, introducido originalmente en un contexto similar por Sinhorn y Knopp (1967). Para esta parte seguiremos el trabajo de Benamou et al. (2014).

Recordamos que el problema de Kantorovich regularizado en el caso discreto puede formularse como

$$P^*_{\gamma,\epsilon} \quad = \quad \underset{P_\gamma \in \Pi(\mu,\nu)}{\arg\min} \langle C, P_\gamma \rangle_F + \epsilon H(\gamma) \qquad (KPD_\epsilon),$$

donde C es la matriz de costo, P_γ es la matriz que representa la medida discreta γ y H es la función de entropía.

En el caso discreto, la divergencia de Kullback-Leibler entre $\gamma \in \mathbb{R}_+^{N\times N}$ y $\xi \in \mathbb{R}_{++}^{N\times N}$ (es decir, $\xi_{i,j} > 0$ para todo i, j) es

$$D_{KL}(\gamma||\xi) = \sum_{i,j=1}^{N} \gamma_{i,j}\left[\log\left(\frac{\gamma_{i,j}}{\xi_{i,j}}\right)\right].$$

Observar que si definimos $\xi = e^{-\frac{C}{\epsilon}}$, donde la exponencial es punto a punto,

entonces

$$\begin{aligned} D_{KL}(\gamma||\xi) &= \sum_{i,j=1}^{N} \gamma_{i,j}\log(\gamma_{i,j}) - \sum_{i,j=1}^{N} \gamma_{i,j}\log(\xi_{i,j}) \\ &= \sum_{i,j=1}^{N} \gamma_{i,j}\log(\gamma_{i,j}) - \sum_{i,j=1}^{N} \gamma_{i,j}\log\left(e^{-\frac{c_{i,j}}{\epsilon}}\right) \\ &= \underbrace{\sum_{i,j=1}^{N} \gamma_{i,j}\log(\gamma_{i,j})}_{H(\gamma)} + \underbrace{\frac{1}{\epsilon}\sum_{i,j=1}^{N} \gamma_{i,j}c_{i,j}}_{\frac{1}{\epsilon}\langle C,P_\gamma\rangle}, \end{aligned}$$

por lo tanto

$$\langle C, P_\gamma\rangle + \epsilon H(\gamma) = \epsilon D_{KL}(\gamma||\xi),$$

lo cual implica que podemos escribir el problema de Kantorovich regularizado como

$$\epsilon \arg\min_{\gamma\in\Pi(\mu,\nu)} D_{KL}(\gamma||\xi) \quad \text{con} \quad \xi = e^{-\frac{C}{\epsilon}}.$$

Por otro lado, si definimos

$$\mathcal{C}_1 = \left\{\gamma \in \mathbb{R}_+^{N\times N} \;\mid\; \gamma\mathbf{1} = \mu\right\} \;\text{ y }\; \mathcal{C}_2 = \left\{\gamma \in \mathbb{R}_+^{N\times N} \;\mid\; \gamma^T\mathbf{1} = \nu\right\},$$

entonces podemos reescribir la restricción $\gamma \in \Pi(\mu,\nu)$ como $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$.

Una noción que necesitamos para el algoritmo de Sinkhorn-Knopp es el de proyección sobre un conjunto convexo:

Definición 3.14 (Proyección KL). *Dado un conjunto convexo $\mathcal{C} \subset \mathbb{R}^{N\times N}$, la proyección KL de ξ sobre $\mathcal{C}$ se define como:*

$$P_{\mathcal{C}}^{KL}(\xi) = \arg\min_{\gamma\in\mathcal{C}} D_{KL}(\gamma||\xi).$$

Con esta notación, la solución al problema de Kantorovich regularizado puede expresarse como

$$P^*_{\gamma,\epsilon} = \epsilon P_{\mathcal{C}}^{KL}(\xi) \;\text{ con }\; \xi = e^{-\frac{C}{\epsilon}} \;\text{ y }\; \mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2.$$

Se puede probar que comenzando en $\gamma^{(0)} = \xi$ y siguiendo la regla de iteración

$$\gamma^{(n)} = P_{\mathcal{C}_n}^{KL}(\gamma^{(n-1)}),$$

se tiene que $\gamma^{(n)}$ converge a la única solución del problema de Kantorovich regularizado, es decir, $\gamma^{(n)} \to P^*_{\gamma,\epsilon}$ cuando $n \to \infty$. Lo interesante, es que las proyecciones sobre $\mathcal{C}_1$ y $\mathcal{C}_2$ tiene una forma explicita que dependen de la iteración anterior y de μ y ν. Este resultado es la proposición 1 de Benamou et al. (2014).

Proposición 3.15. *Dado $\overline{\gamma} \in \mathbb{R}^{N\times N}_+$ tenemos que las proyecciones KL sobre $\mathcal{C}_1$ y $\mathcal{C}_2$ son*

$$P^{KL}_{\mathcal{C}_1}(\overline{\gamma}) = diag\Big(\mu \oslash (\overline{\gamma}\mathbf{1})\Big)\overline{\gamma} \quad y \quad P^{KL}_{\mathcal{C}_2}(\overline{\gamma}) = \overline{\gamma} diag\Big(\nu \oslash (\overline{\gamma}^T\mathbf{1})\Big).$$

En lo anterior el símbolo $\oslash$ representa la división punto a punto entre dos vectores (ver capítulo 1).

A partir de lo anterior, la regla de iteración queda

$$\gamma^{(n)} = diag(u^{(n)})\xi diag(v^{(n)}),$$

donde $(u^{(n)}, v^{(n)})$ son vectores en $\mathbb{R}^N \times \mathbb{R}^N$ que satisfacen que $v^{(0)} = \mathbf{1}$ y obedecen a la siguiente recursion

$$u^{(n)} = \mu \oslash \xi v^{(n)} \quad \text{y} \quad v^{(n+1)} = \nu \oslash \xi^T u^{(n)}.$$

Esto permite implementar un algoritmo para hallar la solución al problema de Kantorovich regularizado que solo utilice multiplicaciones de matrices y vectores, con una matriz ξ fija, haciéndolo así altamente eficiente en una computadora. Este algoritmo es conocido como algoritmo de Sinkhorn-Knopp y presentamos el pseudo-código a continuación:

Algorithm 2 Algoritmo Sinkhorn-Knopp

Require: μ, ν, C, ϵ, L
  $v^{(0)} = 1, \xi = \exp(-\frac{C}{\epsilon})$
  for $l \in 1, \ldots, L$ **do**
    $v^{(l)} = \nu \oslash \xi^T u^{(l-1)}$ {Actualizar reescalado derecho}
    $u^{(l)} = \mu \oslash \xi v^{(l)}$ {Actualizar reescalado izquierdo}
  end for
  return $T = \text{diag}(u^{(L)})\xi\text{diag}(v^{(L)})$

En el pseudo-código anterior, L es la cantidad de iteraciones que se quieren realizar, la misma puede ser un número fijo o alguna condición de convergencia.

La figura 3.2 muestra como varia el plan de transporte al aumentar la cantidad de iteraciones l.

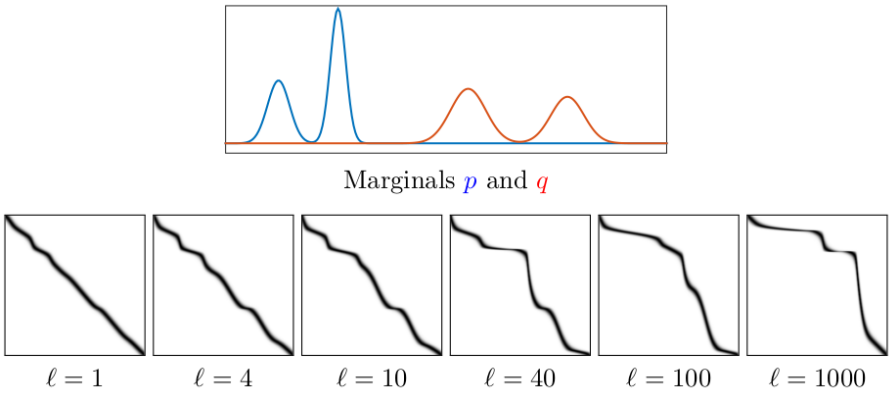


Figura 3.2: La variable $l \in \{1, \ldots, 1000\}$ representa la iteración. Al aumentar la cantidad de iteraciones nos acercamos al plan de transporte óptimo. Figura tomada de Benamou et al. (2014).

3.3. Transporte óptimo en la práctica

En esta sección, realizaremos un repaso detallado de las técnicas más comunes para la aplicación del transporte óptimo en contextos prácticos. Veremos diversos enfoques y metodologías que se han desarrollado para resolver problemas específicos. Este conocimiento será fundamental para el capítulo de experimentos, donde aplicaremos las técnicas discutidas para resolver algunos experimentos y evaluar su rendimiento en escenarios simulados.

Discretización

Dada una muestra $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^n$ de datos, queremos aproximar la densidad subyacente de donde se extrajo la muestra como una medida

$$\hat{\mu} = \sum_{i=1}^{N} a_i \delta_{\mathbf{z}_i},$$

donde $\sum_{i=1}^N a_i = 1$. Tenemos dos formas de encontrar las ponderaciones a_i y los puntos $\mathbf{z}_i$: la discretización Lagrangiana y la discretización Euleriana.

En el primer caso, suponemos que tomamos $\mathbf{z}_i = \mathbf{x}_i$ y fijamos los pesos $a_i = \frac{1}{N}$. En el segundo caso, dividimos el espacio ambiente en una grilla: los

puntos $\mathbf{z}_i$ son el centro de cada celda y a_i corresponde a la cantidad de datos dentro de cada celda. La figura 3.3 muestra ambos métodos. Este segundo enfoque es el que solemos seguir en las aplicaciones, dado que por lo general no hay información adicional para darle más importancia a algunas observaciones que a otras.

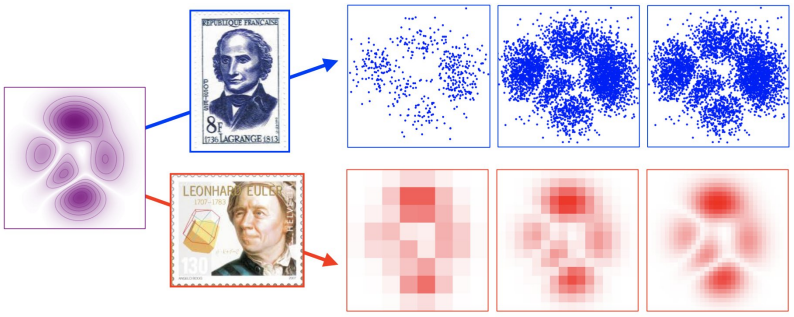


Figura 3.3: A la izquierda una densidad continua. En azul la discretización Lagrangiana y en rojo la discretización Euleriana. Imagen tomada de Gabriel Peyré.

Trabajando con imágenes

Al trabajar con imágenes, es esencial realizar ciertas transformaciones antes de aplicar las técnicas de transporte óptimo. Estas transformaciones permiten representar las imágenes de manera adecuada para los experimentos y análisis subsecuentes. A continuación, se presentan dos escenarios utilizados en los experimentos:

1. Representación de cada imagen como un punto en $[0,1]^n \subset \mathbb{R}^n$: en este escenario, cada imagen se considera como un punto en un espacio de alta dimensión, donde n es el número de características utilizadas para describir la imagen.
2. Representación de cada imagen en el espacio de píxeles: cada imagen se descompone en sus píxeles constituyentes, y cada píxel se representa como un punto en $[0,1] \times [0,1] \times [0,1] \subset \mathbb{R}^3$, donde cada dimensión representa una componente de color.

Estas transformaciones son fundamentales para aplicar de manera efectiva las técnicas de transporte óptimo en el procesamiento y análisis de imágenes. En lo que sigue formalizaremos ambos enfoques.

Transporte óptimo entre colección de imágenes

Comencemos viendo como se trabaja con imágenes en escalas de grises. Una manera natural de representar una imagen en escala de grises de tamaño $w \times h$ es con una matriz $I \in \mathcal{M}(\mathbb{R})_{w\times h}$. En lo anterior w es el ancho (width) mientras que h es la altura (height). Cada elemento de la matriz representa la intensidad de un píxel de la imagen, la cual puede tomar valores entre 0 (negro absoluto) y 255 (blanco absoluto). La figura 3.4 muestra un ejemplo sobre una imagen de una cara.

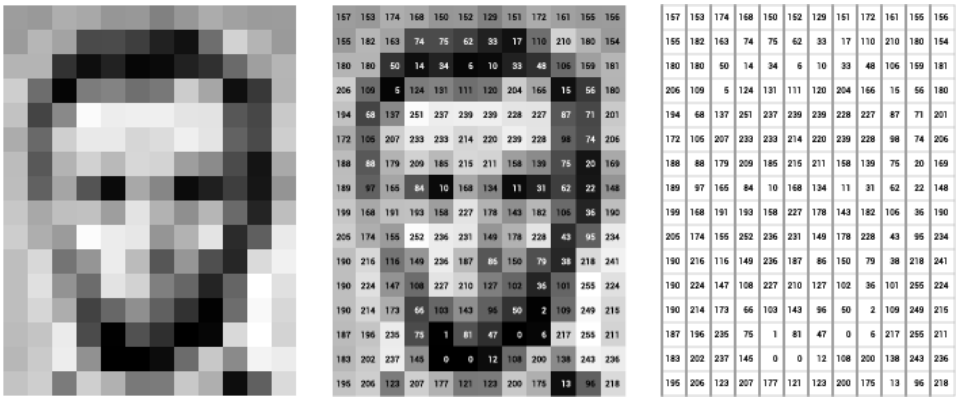


Figura 3.4: Representación de una imagen en escala de grises en una computadora. Imagen tomada de ai.stanford.edu.

Es usual normalizar los valores de los píxeles y definir $\tilde{I} = \frac{1}{255} I$ para que la intensidad varíe entre 0 y 1. Por otro lado, para trabajar con transporte óptimo necesitamos vectores en $\mathbb{R}^n$ por lo que suele transformarse la matriz $\hat{I}$ vectorizandola en un vector $\mathbf{x}_{\hat{I}}$ de la siguiente forma:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \longrightarrow \text{vectorizar} \longrightarrow \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \\ a_{31} \\ a_{32} \\ a_{33} \end{bmatrix}$$

Esta operación se denomina vectorización (flattening en ingles). Luego de

representar cada imagen I como un vector en $\mathbb{R}^{w\times h}$ utilizamos la discretización Lagragiana, es decir, suponemos una distribución uniforme en los datos. Explícitamente, si tenemos un conjunto de imágenes $\{I_k\}_{k=1}^K$ entonces aproximamos la distribución de probabilidad que generó los datos mediante la medida discreta

$$\hat{\mu} = \sum_{k=1}^{K} \frac{1}{K} \delta_{\mathbf{x}_{\hat{I}_k}} \in \mathcal{P}([0,1]^{w\times h}).$$

El siguiente paso es extender las ideas anteriores para el contexto de imágenes con color. Una representación usual de imágenes a color es el modelo RGB (Red, Green, Blue). En este modelo, los colores se crean combinando diferentes intensidades de los colores primarios de la luz: rojo, verde y azul. La representación en una computadora es con una matriz para cada color, teniendo así un objeto con dimensiones $w \times h \times c$ donde c es la cantidad de canales (channels). En la literatura estos objetos suelen llamarlos cubos. La figura 3.5 muestra un esquema de la representación. Fijado un canal (por ejemplo el rojo) cada elemento de la matriz representa la intensidad de ese color en el píxel dado. Por otro lado, fijado un píxel tenemos ahora tres números que lo representan, cada uno representando la intensidad de rojo, azul y verde de ese píxel.

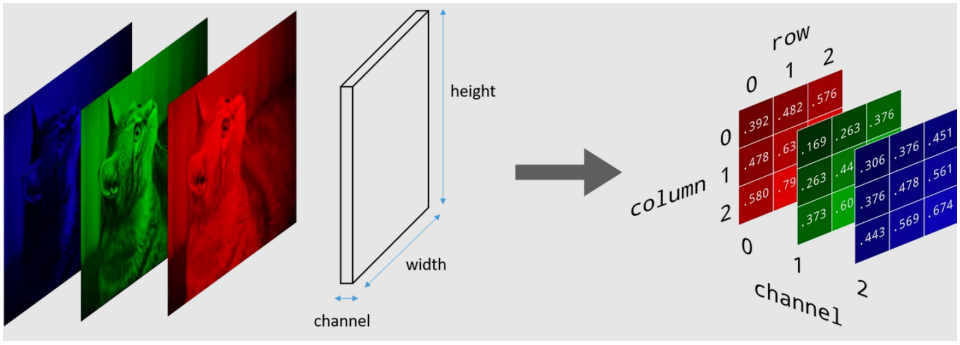
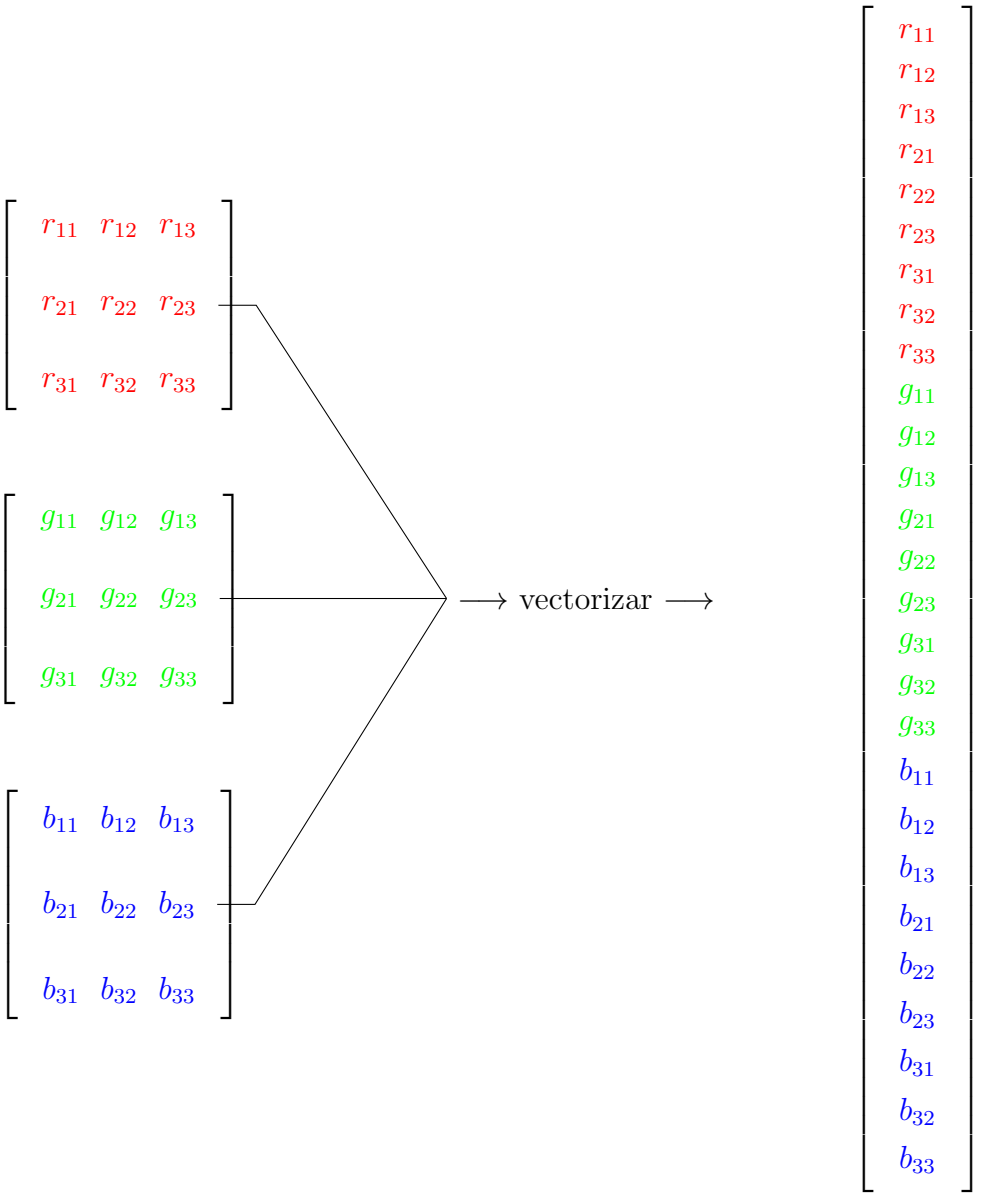


Figura 3.5: Representación de una imagen RGB en una computadora. Imagen tomada de towardsdatascience.com.

En este escenario podemos vectorizar cada canal, obteniendo así 3 vectores $\mathbf{x}_{I_R}, \mathbf{x}_{I_G}, \mathbf{x}_{I_B} \in \mathbb{R}^{w\times h}$ y luego concatenarlos para generar un único vector en $\mathbf{x}_I \in \mathbb{R}^{w\times h\times 3}$. Al igual que con las imágenes en escalas de grises, si $\{I_k\}_{k=1}^K$ es una colección de imágenes entonces consideramos como aproximación de la distribución de probabilidad a la medida discreta

$$\hat{\mu} = \sum_{k=1}^{K} \frac{1}{K} \delta_{\mathbf{x}_{\hat{I}_k}} \in \mathcal{P}([0,1]^{w\times h\times 3})$$



En este contexto, si tenemos dos colecciones de imágenes RGB, digamos $\{I_k\}_{k=1}^K$ y $\{J_l\}_{l=1}^L$, el primer paso es llevar todas las imágenes a una misma dimensión $w \times h \times 3$. Esto se puede hacer por ejemplo recortando las imágenes o utilizando zero-padding (agregar ceros/negro en los extremos hasta conseguir la dimensión deseada). Esto es necesario para que, al momento de vectorizar, todas las imágenes terminen teniendo la misma dimensión. Luego de vectorizar ambas colecciones tendremos:

$$\hat{\mu} = \sum_{k=1}^{K} \frac{1}{K} \delta_{\mathbf{x}_{i_k}} \quad \text{y} \quad \hat{\nu} = \sum_{l=1}^{L} \frac{1}{L} \delta_{\mathbf{x}_{j_l}} \subset \mathcal{P}([0,1]^{w \times h \times 3}).$$

Luego la matriz de costo tendrá dimensiones $K \times L$ donde la entrada c_{kl} representa el costo de transportar la imagen k-ésima (vectorizada) hasta la imagen l-ésima. En los experimentos utilizamos la distancia euclídea como función de costo, así

$$C = ((c_{kl}))_{kl} \in \mathcal{M}(\mathbb{R})_{K \times L} \quad \text{donde} \quad c_{kl} = ||\mathbf{x}_{\hat{I}_k} - \mathbf{x}_{\hat{J}_l}||_2.$$

Es importante remarcar que esta no es la única distancia, y tampoco tiene por que ser la mejor, dado que la elección de la distancia esta fuertemente condicionada por la aplicación. En nuestro caso vamos a utilizar la distancia euclídea por la rapidez computacional al calcular el transporte óptimo.

Finalmente, el plan de transporte también será una matriz $\gamma = ((\gamma_{kl}))_{kl}$ de tamaño $K \times L$, donde la entrada γ_{kl} representa la masa a transportar de la imagen $\mathbf{x}_{\hat{I}_k}$ hacia la imagen $\mathbf{x}_{\hat{J}_l}$.

Transferencia de color entre dos imágenes

En la transferencia de color, el objetivo es transportar los colores de una imagen a otra. El uso del transporte óptimo en la transferencia de color es especialmente eficaz porque garantiza una correspondencia globalmente óptima de los colores, preservando las estructuras y relaciones entre colores originales, lo que da como resultado una imagen visualmente coherente y atractiva. Para resolver la transferencia de color utilizando el transporte óptimo, se sigue generalmente este procedimiento:

1. Las imágenes se representan en un espacio de color adecuado, por ejemplo RGB, donde cada píxel tiene un valor de color.
2. Se calculan las distribuciones de probabilidad de los colores para la imagen destino (la que se va a transformar) y la imagen de referencia.
3. Se calcula y aplica el transporte óptimo entre ambas distribuciones a los píxeles de la imagen destino para que sus colores coincidan con los de la imagen de referencia.

A continuación profundizaremos en cada etapa de este proceso. Como vimos anteriormente, una imagen RGB se puede representar como un cubo de dimensiones $w \times h \times 3$. A diferencia del enfoque anterior, al trabajar en la transferencia de color en lugar de representar cada imagen I como un vector en $\mathbb{R}^{w \times h \times 3}$ vamos a representarla como su colección de píxeles. Así en lugar

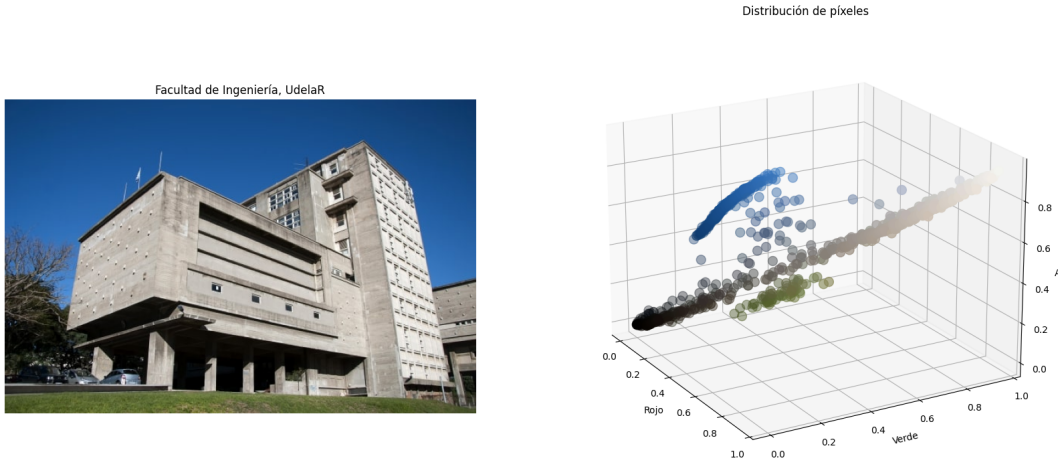


Figura 3.6: A la izquierda una fotografía de la Facultad de Ingeniería, UdelaR. A la derecha su representación en el espacio de píxeles. Este último consiste en representar cada píxel de la imagen como una terna de números (r_i^I, g_i^I, b_i^I) representando la intensidad del mismo en cada uno de los canales rojo, verde y azul respectivamente.

de pensar a una imagen como 3 matrices (cada una representando un color) podemos pensar a la imagen como una colección de píxeles, cada uno determinado por 3 valores: su intensidad de rojo, verde y azul. Luego, volvemos a utilizar la discretización Lagragiana, es decir, le damos el mismo peso a cada píxel, por lo tanto, la distribución de probabilidad estimada de la imagen será

$$\hat{\mu}_I = \sum_{i=1}^{w \times h} \frac{1}{w \times h} \delta_{\mathbf{p}_i^I} \in \mathcal{P}([0,1]^3),$$

donde $\mathbf{p}_i^I = (r_i^I, g_i^I, b_i^I) \in \mathbb{R}^3$ es el píxel $i-$ésimo de la imagen.

La figura 3.6 muestra una imagen de la Facultad de Ingeniería (FING) y su correspondiente distribución de píxeles. Es esperable que la mayoría de píxeles estén en posiciones que representan los grises, dado que la mayor parte de la fotografía está conformada por el edificio principal y es de este color. Por otro lado esperamos ver una colección de píxeles en azul (el cielo) y otra colección de píxeles en verde (el pasto).

En este contexto, tenemos solo dos imágenes I y J, de tamaño $w \times h \times 3$ y $w' \times h' \times 3$ respectivamente. Aquí no es necesario que las dos imágenes tengan la misma altura y anchura. Luego, realizando el proceso descrito anteriormente tenemos

$$\hat{\mu}_I = \sum_{i=1}^{w\times h} \frac{1}{w \times h}\delta_{\mathbf{p}_i^I} \quad \text{y} \quad \hat{\nu}_J = \sum_{j=1}^{w'\times h'} \frac{1}{w' \times h'}\delta_{\mathbf{p}_j^J} \in \mathcal{P}([0,1]^3).$$

Estamos representando la imagen I con $w \times h$ puntos y la imagen J con $w' \times h'$ puntos por lo tanto la matriz de costo tendrá dimensiones $(w \times h) \times (w' \times h')$. En la práctica solemos usar la distancia euclídea, por lo tanto

$$C = (c_{ij})_{ij} = ||\mathbf{p}_i^I - \mathbf{p}_j^J||_2,$$

es la distancia euclídea entre el píxel i-ésimo de una imagen y el píxel j-ésimo de la otra. Finalmente, el transporte óptimo tendrá el mismo tamaño que la matriz de costo, es decir, $(w \times h) \times (w' \times h')$.

El procedimiento anterior puede llevar a problemas de rendimientos, dado que el costo computacional de resolver el problema de transporte óptimo depende de la cantidad de datos y en este caso tenemos que transportar $w \times h$ datos (los píxeles). En la actualidad es normal que la mayoría de los teléfonos celulares modernos tengan sensores que permiten capturar fotografía con decenas de magapixeles, por ejemplo 12 píxeles, esto se traduce es 12 millones de píxeles para transportar. Por esta razón, se suelen elegir una cantidad menor de píxeles al azar (por ejemplo 1000) y transportar únicamente estos. La desventaja de este procedimiento es que implica una disminución en la calidad de la imagen destino, por lo tanto se tiene un compromiso entre tiempo de cómputo y calidad del resultado.

3.4. Transporte óptimo y Machine Learning

Para concluir este capítulo, haremos un repaso de como el transporte óptimo ha encontrado aplicaciones en varios campos dentro de machine learning, demostrando su versatilidad y eficacia por ejemplo en la solución de problemas de adaptación de dominio, generación de datos sintéticos. A continuación, se presentan algunas de las aplicaciones más relevantes.

1. **Generación de Datos Sintéticos**
   El transporte óptimo se utiliza en la generación de datos sintéticos para mejorar la diversidad y calidad de los conjuntos de datos. Al transformar distribuciones conocidas a nuevas distribuciones deseadas, es posible

generar datos sintéticos que preserven las propiedades estadísticas esenciales del conjunto de datos original. Esto es particularmente útil en tareas como la ampliación de conjuntos de datos para aprendizaje profundo y la generación de imágenes realistas en aplicaciones de visión por computadora. Algunos ejemplos en esta linea de trabajos son: Wasserstein GAN (Arjovsky et al. 2017) o "Learning Generative Models with Sinkhorn Divergences" (Genevay et al. 2017).

2. **Transferencia de Estilo en Imágenes** Utilizando distancias de transporte óptimo, es posible transformar el estilo de una imagen manteniendo su contenido esencial. Esta técnica ha sido empleada en la creación de efectos artísticos y en la adaptación de imágenes entre diferentes estilos visuales. Un ejemplo es el artículo "Style Transfer by Relaxed Optimal Transport and Self-Similarity" de Kolkin et al. (2019) donde los autores usan una versión relajada de la distancia de Wasserstein.

3. **Análisis de Datos Temporales**
   En el análisis de series temporales, el transporte óptimo ofrece una forma efectiva de comparar y alinear secuencias temporales. Esto es útil en aplicaciones como la detección de anomalías, el análisis de patrones y la previsión temporal, donde es fundamental medir la similitud entre series temporales de manera robusta. Un ejemplo es el artículo "OTW: Optimal transport warping for time series" de Latorre et al. (2023).

4. **Adaptación de Dominio**
   En machine learning, la adaptación de dominio es crucial cuando los datos de entrenamiento (dominio de origen) y los datos de prueba (dominio de destino) provienen de distribuciones diferentes. El transporte óptimo puede alinear estas distribuciones, mejorando la precisión del modelo al minimizar la discrepancia entre ellas. Un ejemplo destacado es el uso del transporte óptimo regularizado para ajustar las representaciones de características entre dominios, permitiendo que un modelo entrenado en un dominio sea eficaz en otro. En esto nos concentramos en el próximo capítulo.

Estas son solo algunas de las aplicaciones que el transporte óptimo tiene dentro de machine learning. El lector interesado en incursionar en estas y otras aplicaciones puede consultar Montesuma et al. (2023).

Por último, queremos destacar el impacto significativo que está teniendo el transporte óptimo en la comunidad de machine learning. Un ejemplo claro de esto es la edición 2023 de la conferencia NeurIPS (Neural Information Processing Systems), donde se dedicó un workshop específico al transporte óptimo en machine learning. NeurIPS es una de las conferencias más prestigiosas en el campo del machine learning, donde fueron presentados en distintas ediciones de la misma trabajos de mucha relevancia como: Word2Vec (Mikolov et al. 2013), InstructGPT (Ouyang et al. 2022) y Diffusion Models (Saharia et al. 2022).

Capítulo 4

Adaptación de Dominio

En este capítulo exploraremos la adaptación de dominio, una técnica de aprendizaje automático la cual permite aplicar modelos entrenados en un dominio fuente (source) a otro dominio objetivo (target) con características diferentes. La adaptación de dominio se ha convertido en una herramienta esencial para abordar problemas donde la recolección de datos etiquetados es costosa o impracticable, o cuando las diferencias entre los conjuntos de datos de fuente y objetivo pueden afectar significativamente el rendimiento del modelo considerado inicialmente.

En la primera sección discutiremos algunas razones por las cuales la adaptación de dominio es necesaria y los beneficios que ofrece en diversos campos de aplicación. Luego, proporcionaremos las definiciones clave y una taxonomía detallada de los distintos tipos de adaptación de dominio. Será importante entender esta clasificación para poder comprender en cual contexto es posible aplicar el transporte óptimo. Nos centraremos en la adaptación de dominio no supervisado, donde a diferencia del dominio original, no se dispone de etiquetas en el dominio objetivo. Abordaremos este problema utilizando el transporte óptimo. Discutiremos los desafíos específicos de trabajar sin datos etiquetados y cómo superar estos obstáculos con enfoques innovadores.

4.1. Motivación

En cualquier método de machine learning una hipótesis fundamental es que la distribución de los datos de entrenamiento (train set) y los datos de prueba (test set) provienen de la misma distribución. Sin embargo, esta suposición

no siempre se cumple en la práctica. Las diferencias entre las distribuciones de datos pueden surgir por diversas razones, como cambios en las condiciones ambientales, variaciones en los dispositivos de adquisición o diferencias en las características demográficas. Es en esta situación en donde la adaptación de dominio cobra relevancia, ya que aborda la disparidad entre los conjuntos de datos de entrenamiento y prueba.

Las discrepancias en la distribución de datos (también conocidas como 'drift') se deben a varias razones y dependen de la aplicación. En Computer Vision, el drift ocurre al cambiar las condiciones de iluminación, dispositivos de adquisición, o considerando la presencia o ausencia de fondo. En el procesamiento del lenguaje (speech processing), aprender de un locutor e intentar implementar una aplicación dirigida a un público amplio también puede verse obstaculizada por las diferencias en el ruido de fondo, el tono o el género del vocero.

En el artículo "Learning under Concept Drift: A Review" Lu et al. (2020) estudian el concepto de drift en varias situaciones así como metodologías para evitar su efecto. Un ejemplo más específico puede ser el artículo "Data drift in medical machine learning: implications and potential remedies" Sahiner et al. (2023) donde se estudia los efectos de drift en aplicaciones médicas. Estos problemas motivaron la búsqueda de técnicas que permitan entrenar un modelo en un dominio diferente al dominio utilizado en producción, bajo la suposición de que la tarea a realizar es la misma.

4.2. Definiciones básicas y taxonomía

En esta sección, presentamos una categorización de los distintos tipos de adaptación de dominio descritos en la literatura. La misma se basa en el trabajo de Wang y Deng (2018), en el que proporcionan un marco exhaustivo para categorizar las diversas estrategias y enfoques utilizados en la adaptación de dominio. Primero, revisaremos las definiciones básicas necesarias para comprender los conceptos clave y establecer una base sólida para la discusión. A continuación, exploraremos las diferentes categorías de adaptación de dominio, destacando sus características principales. Esta clasificación incluye enfoques supervisados, no supervisados, y semisupervisados, así como técnicas específicas como el alineamiento de distribuciones, el uso de redes adversarias y

métodos basados en la representación. Antes de continuar queremos remarcar que en esta sección utilizaremos una notación levemente diferente a la usada en las secciones anteriores. Antes utilizamos $\mathcal{X}$ e $\mathcal{Y}$ para los conjuntos source y target, pero en esta sección tanto el conjunto source como target tendrán una parte de observaciones $\mathcal{X}$ y una parte de etiquetas $\mathcal{Y}$. Así por ejemplo, el conjunto source es de la forma $\mathcal{X}^s \times \mathcal{Y}^s$. Comencemos con algunas definiciones básicas.

Definición 4.1 (Dominio). *Un dominio $\mathcal{D}$ es un par $(\mathcal{X}, P(X))$, donde $\mathcal{X}$ es en un espacio de características y $P(X)$ es una distribución de probabilidad, con $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X}$.*

Para fijar ideas, podemos suponer que queremos entrenar un modelo sobre imágenes de animales, por ejemplo: gatos, perros, caballos, leones y cebras. En este contexto el espacio de características $\mathcal{X}$ podrían ser los histogramas de las imágenes o descriptores de textura, los cuales son métricas diseñadas para cuantificar la textura percibida en una imagen. Por otro lado, el conjunto $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ es una colección de estas imágenes y $P(X)$ puede representar la probabilidad de que una imagen dada sea de un animal, por ejemplo, el conjunto de datos podría estar compuesto por un 20 % de imágenes de gatos, un 30 % de imágenes de perros, un 15 % de imágenes de caballos, un 10 % de imágenes de leones, y un 25 % de imágenes de cebras. Dado un dominio $\mathcal{D} = \{\mathcal{X}, P(X)\}$, necesitamos definir una tarea a realizar.

Definición 4.2 (Tarea). *Una tarea $\mathcal{T}$ es un par $(\mathcal{Y}, f(\cdot))$, donde $\mathcal{Y}$ es un espacio de característica y $f : \mathcal{X} \to \mathcal{Y}$ es una función predictiva.*

Siguiendo con el ejemplo de los animales, tendríamos que $\mathcal{Y} = \{\text{Gato, Perro, Caballo, León, Cebra}\}$ mientras que $f(\cdot)$ podría ser un modelo de clasificación como un Support Vector Machine (SVM) o una red neuronal convolucional (CNN).

Si suponemos que $Y \subset \mathcal{Y}$ es un conjunto de etiquetas entonces los datos de entrenamiento del modelo son de la forma $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, por ejemplo $(\mathbf{x}_i, \text{León})$. La función $f(\cdot)$ puede interpretarse como la probabilidad condicional $P(Y|X)$, es decir, dado el dato $\mathbf{x}_i$, la función $f(\mathbf{x}_i)$ nos dice la probabilidad que dicho dato pertenezca a la clase y_i. El objetivo fundamental de cualquier modelo de aprendizaje automático supervisado es estimar (o aprender) la función $f(\cdot)$ a partir de datos etiquetados $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$. Este proceso

implica que, dado un conjunto de datos $\mathbf{x}_i$, que representan las características observables, y las correspondientes etiquetas y_i, que representan las salidas deseadas o categorías, el modelo busca identificar una relación subyacente entre estas dos variables.

En el escenario de la adaptación de dominio tenemos dos dominios: el dominio source (o dominio de entrenamiento) $\mathcal{D}^s = \{\mathcal{X}^s, P(X)^s\}$ con suficientes datos etiquetados y un dominio target $\mathcal{D}^t = \{\mathcal{X}^t, P(X)^t\}$ donde hay pocas o ninguna etiqueta. Podemos descomponer el dominio target como $\mathcal{D}^t = \mathcal{D}^{tl} \cup \mathcal{D}^{tu}$, donde $\mathcal{D}^{tl}$ y $\mathcal{D}^{tu}$ son la parte etiquetada y no etiquetada, respectivamente. Cada dominio tiene una tarea asociada, $\mathcal{T}^s = \{\mathcal{Y}^s, f^s(\cdot)\}$ y $\mathcal{T}^t = \{\mathcal{Y}^t, f^t(\cdot)\}$. Además $f^s(\cdot)$ puede estimarse a partir de los datos etiquetados del conjunto source $\{\mathbf{x}_i^s, y_i^s\}$ mientras que $f^t(\cdot)$ puede aprenderse a partir de la parte etiquetada del conjunto target $\{\mathbf{x}_i^{tl}, y_i^{tl}\}$.

Como mencionamos en la introducción de este capítulo, el aprendizaje automático tradicional asume que los dominios son iguales, $\mathcal{D}^s = \mathcal{D}^t$, y que las tareas a realizar son la misma, es decir, $\mathcal{T}^s = \mathcal{T}^t$. En la adaptación de dominio suponemos que las tareas son las mismas ($\mathcal{T}^s = \mathcal{T}^t$), pero que los dominios son distintos $\mathcal{D}^s \neq \mathcal{D}^t$. Esta diferencia entre los dominios puede deberse a dos razones, lo cual conduce a dos categorías de la adaptación de dominio: diferencias en la distribución $P(X)^s \neq P(X)^t$ al cual llamaremos adaptación de dominio homogénea, o diferencias en el espacio de características $\mathcal{X}^s \neq \mathcal{X}^t$, la cual denominaremos adaptación de dominio heterogénea.

Por otro lado, cada una de estas categorías puede subdividirse en 3 subcategorias dependiendo de la cantidad de datos etiquetados en el dominio target: adaptación de dominio supervisado (SDA por Supervised Domain Adaptation), adaptación de dominio semi-supervisado (SSDA por Semi Supervised Domain Adaptation) y adaptación de dominio no supervisado (UDA por Unsupervised Domain Adaptation), siendo esta última donde utilizaremos el transporte óptimo. La tabla 4.1 muestra las particularidades de cada una de estas 3 subcategorias.

Por último, existe otro tipo de adaptación de dominio que no entra en ninguno de los mencionados anteriormente: la adaptación de dominio multifuente. En esta última, se disponen múltiples dominios fuentes y se busca adaptar el

Tipo	Etiquetas en Target	Aplicación
Supervisada (SDA)	Todos los datos etiquetados	Útil cuando es posible etiquetar una gran cantidad de datos en el dominio objetivo. Por ejemplo, un sistema de traducción automática que se adapta a un nuevo dialecto con algunos ejemplos etiquetados.
Semi-Supervisada (SSDA)	Algunos datos etiquetados	Situaciones donde se puede etiquetar una parte de los datos en el dominio objetivo, pero la mayoría de los datos siguen sin etiquetar. Por ejemplo, en la clasificación de spam en correos electrónicos donde algunos correos están etiquetados y otros no.
No Supervisada (UDA)	Ningún dato etiquetado	Común cuando obtener datos etiquetados en el nuevo dominio es costoso o impracticable. Un ejemplo es adaptar un modelo de reconocimiento de objetos entrenado en imágenes de alta calidad a imágenes de baja resolución o diferentes condiciones de iluminación.

Tabla 4.1: Clasificación de los tipos de adaptación de dominio según la cantidad de datos etiquetados en el dominio target.

modelo a un único dominio objetivo, un ejemplo sobre esta linea de trabajo es el artículo de Redko, Courty y Tuia (Redko et al. 2019).

En este trabajo nos concentraremos principalmente en la adaptación de dominio homogénea no supervisada (UDA), y en la aplicación del transporte óptimo en este contexto, donde tendremos como referencia principal el artículo "Optimal Transport for Domain Adaptation" de Courty et al. (2016).

La adaptación de dominio se enmarca en una teoría más general, llamada transfer learning, la cual consiste en transferir conocimiento de una modelización source a una modelización target y donde los dominios y las tareas pueden

ser o no diferentes. El lector interesado puede consultar los artículos de Pan y Yang (2010) y Cao et al. (2023).

4.3. Adaptación de Dominio no Supervisado

En esta sección presentaremos la adaptación de dominio no supervisada, basándonos principalmente en el trabajo de Courty et al. (2016). Exploraremos las técnicas y metodologías clave que permiten transferir conocimientos entre dominios con distribuciones diferentes sin necesidad de etiquetas en el dominio objetivo.

En esta sección supondremos que $\mathcal{X}$ es un espacio de probabilidad en $\mathbb{R}^n$ y llamaremos $\mathcal{Y}$ al conjunto de posibles etiquetas. Continuamos con algunas definiciones necesarias. Comencemos definiendo el conjunto en donde entrenaremos un modelo:

Definición 4.3 (Conjunto Source). *Dado un dominio $\mathcal{X}^s$ y un espacio de etiquetas $\mathcal{Y}$, llamaremos conjunto source al conjunto de entrenamiento $\boldsymbol{X}^s \times Y^s \subset \mathcal{X}^s \times \mathcal{Y}$, en donde $\boldsymbol{X}^s = \{\boldsymbol{x}_i^s\}_{i=1}^{n_s}$ es un conjunto de datos e $Y^s = \{y_i^s\}_{i=1}^{n_s}$ con $y_i^s \in \mathcal{Y}$ es el conjunto de etiquetas asociados a esos datos.*

Por otro lado, el conjunto test se define como:

Definición 4.4 (Conjunto Target). *Dado un dominio $\mathcal{X}^t$, llamaremos conjunto target al conjunto de predicción $\boldsymbol{X}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{n_t} \subset \mathcal{X}^t$, donde las etiquetas son desconocidas.*

En el paradigma tradicional del aprendizaje automático, asumimos la hipótesis de que tanto $\mathbf{X}^s$ como $\mathbf{X}^t$ provienen de una misma variable aleatoria y por lo tanto tienen igual distribución de probabilidad, por lo tanto podemos entrenar un modelo sobre los pares $(\mathbf{x}_i^s, y_i^s)$ para luego inferir las etiquetas de $\mathbf{x}_i^t$.

Sin embargo, en el paradigma de adaptación de dominio, partimos de la suposición de que los $\mathbf{X}^s$ y $\mathbf{X}^t$ provienen de variables aleatorias diferentes y por lo tanto de distribuciones diferentes.

Definición 4.5. *Denotaremos μ_s y μ_t a las distribuciones de $\boldsymbol{P}^s$.*

En la mayoría de los problemas de de Adaptación de Dominio se suele asumir una de las dos hipótesis siguiente:

1. **Desequilibrio de clases:** la distribución de las etiquetas es distinta pero las probabilidades condicionales de las muestras son iguales $\mathbf{P}^s(\mathbf{x}^s|y) = \mathbf{P}^t(\mathbf{x}^t|y)$.
2. ***Covariate Shift*:** las probabilidades condicionales de las etiquetas con respecto a los datos son iguales $\mathbf{P}^s(y|\mathbf{x}^s) = \mathbf{P}^t(y|\mathbf{x}^t)$ pero las distribuciones de los datos son distintas $\mathbf{P}^s(\mathbf{x}^s) \neq \mathbf{P}^t(\mathbf{x}^t)$.

Asumiremos que el shift sobre las distribuciones de probabilidad ocurre a través de una función desconocida $T : \mathcal{X}^s \to \mathcal{X}^t$ y que ésta preserva la distribución condicional, es decir

$$\mathbf{P}^s(y|\mathbf{x}^s) = \mathbf{P}^t(y|T(\mathbf{x}^s)).$$

Desde el punto de vista del transporte óptimo, el mapa T no es otra cosa que un transporte entre μ_s y μ_t, es decir, $T_\sharp \mu_s = \mu_t$ (figura 4.1). A partir de esta premisa, tenemos un enfoque sistemático para resolver el problema del cambio de distribución:

1. Estimar μ_s y μ_t a partir de $\mathbf{X}^s$ y $\mathbf{X}^t$.
2. Encontrar un transporte óptimo T entre μ_s y μ_t.
3. Utilizar T para transportar los datos etiquetados $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ y entrenar un clasificador para éstos.

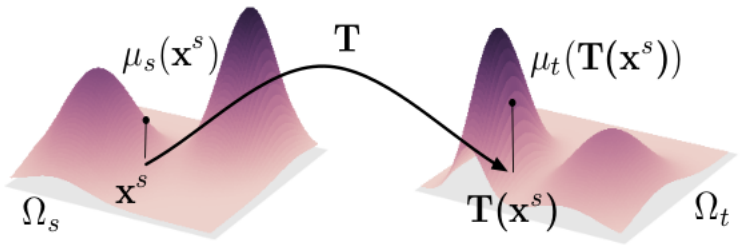


Figura 4.1: Ilustración del push-forward T. Figurada tomada del artículo de Courty, Flamary, Tuia y Rakotomamonjy (2016)

La figura 4.2 ilustra este procedimiento de adaptación de dominio utilizando el plan de transporte obtenido al resolver el problema de Kantorovich.

Como vimos anteriormente, cuando trabajamos en un problema de transporte óptimo solemos resolver el problema de Kantorovich en lugar del de

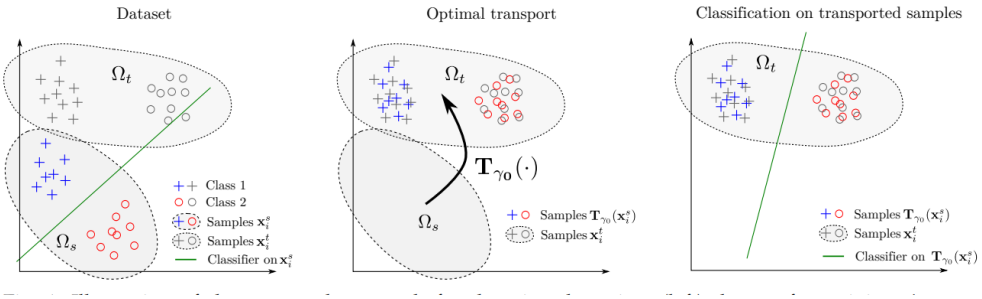


Figura 4.2: Diagrama de la adaptación de dominio. Figurada tomada del trabajo de Courty, Flamary, Tuia y Rakotomamonjy (2016)

Monge. El problema de Kantorovich en este contexto se plantea como en la definición 2.10:

Definición 4.6 (Problema de Kantorovich para la Adaptación de Dominio). *Sean μ_s y μ_t dos distribuciones correspondiente a dos conjuntos de datos $\boldsymbol{X}^s$ y $\boldsymbol{X}^t$ respectivamente, $\gamma \in \Pi(\mu_s, \mu_t)$ un plan de transporte y $c : \mathcal{X}^s \times \mathcal{X}^t \to \mathbb{R}^+$ una función de costo. El problema de Kantorovich es el problema de optimización*

$$\gamma_0 = \underset{\gamma \in \Pi(\mu_s, \mu_t)}{\arg\min} \left\{ \int_{\mathcal{X}^s \times \mathcal{X}^t} c(\mathbf{x}^s, \mathbf{x}^t)\, d\gamma(\mathbf{x}^s, \mathbf{x}^t) \right\}$$

En las aplicaciones trabajaremos con datos discretos $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{n_s}$ y $\mathbf{X}^t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$, por lo que podemos representar la función de costo como una matriz $\mathbf{C} = (c(\mathbf{x}_i^s, \mathbf{x}_j^t))_{ij}$ de tamaño $n_s \times n_t$. En el caso discreto la medida γ también se puede representar mediante una matriz $P_\gamma = (\gamma_{ij})_{ij}$. Con estas notaciones, el problema de Kantorovich se escribe como:

$$\gamma_0 = \underset{\gamma \in \Pi(\mu_s, \mu_t)}{\arg\min} \sum_{i,j} \gamma_{ij} C_{ij} = \underset{\gamma \in \Pi(\mu_s, \mu_t)}{\arg\min} \langle \gamma, \mathbf{C} \rangle_F,$$

donde $\langle \cdot \rangle$ es el producto de Frobenius (ver sección 1). Recordamos que la sección 3.1 contiene una detallada explicación sobre cómo discretizar el problema de Monge y el de Kantorovich.

Una vez que logramos encontrar un plan de transporte óptimo γ_0 necesitamos una manera de "transportar" $\mathbf{X}^s$ a $\mathbf{X}^t$. Siguiendo Courty et al. (2016) nos restringiremos al caso en el cual la función de costo es el cuadrado de la distancia euclídea $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$. Usando la distancia de Wasserstein podemos

interpolar μ_s y μ_t siguiendo la estrategia vista en 2.21. En este contexto, la interpolación de Wasserstein queda

$$\hat{\mu}_\tau = \arg\min_{\mu} (1-\tau) W_2(\mu_s, \mu)^2 + \tau W_2(\mu_t, \mu)^2,$$

donde $\tau \in [0, 1]$.

En el libro de Cédric Villani (2009) se muestra que para este costo, la interpolación resulta en

$$\hat{\mu}_\tau = \sum_{i,j} \gamma_0(i,j) \delta_{(1-\tau)\mathbf{x}_i^s + \tau \mathbf{x}_j^t}.$$

Como nuestro interés consiste en llevar los datos desde el dominio $\mathbf{X}^s$ al dominio $\mathbf{X}^t$, solo nos interesa lo que resulta cuando $\tau = 1$, es decir

$$\hat{\mu} = \sum_j \hat{p}_j \delta_{\mathbf{x}_j^t} \quad \text{donde} \quad \hat{p}_j = \sum_i \gamma_0(i,j).$$

Lo anterior nos dice que los datos transportados por el plan γ_0 son una combinación lineal de $\{\mathbf{x}_1^t, \ldots, \mathbf{x}_{n_t}^t\}$ y por lo tanto $\hat{\mu}$ está soportados en $\{\mathbf{x}_j^t\}_j$.

A continuación presentamos un ejemplo para explicitar la observación anterior.

Ejemplo 4.7. *Sean* $\mathbf{X}^s = \{(1,1),(1,3),(3,3),(4,1),(6,3)\}$ *y* $\mathbf{X}^t = \{(2,2),(5,2)\}$ *dos conjuntos de puntos. Supongamos que la medida* μ_s *y* μ_t *son medidas uniformes sobre* $\mathbf{X}^s$ *y* $\mathbf{X}^t$*, es decir,*

$$\mu_s = \sum_{i=1}^{5} \frac{1}{5} \delta_{\mathbf{x}_i^s} \quad \mu_t = \sum_{j=1}^{2} \frac{1}{2} \delta_{\mathbf{x}_j^t}.$$

Utilizaremos como función de costo el cuadrado de la distancia euclídea $c(\mathbf{x}_i^s, \mathbf{x}_j^t) = \|\mathbf{x}_i^s - \mathbf{x}_j^t\|_2^2$*, con lo cual la matriz de costo* C *es*

$$C = \begin{pmatrix} 2 & 17 \\ 2 & 17 \\ 2 & 5 \\ 5 & 2 \\ 17 & 2 \end{pmatrix}$$

En este caso es sencillo resolver visualmente el problema, dado que el costo

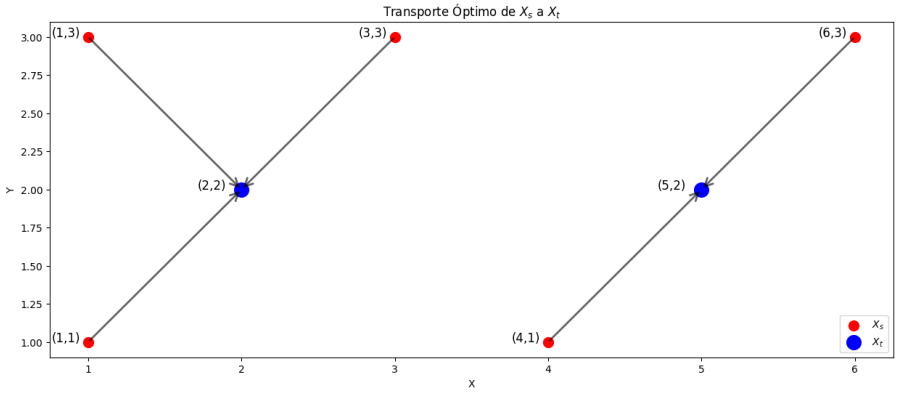


Figura 4.3: En rojo los datos $\mathbf{X}^s = \{\mathbf{x}_1^s, \mathbf{x}_2^s, \mathbf{x}_3^s, \mathbf{x}_4^s, \mathbf{x}_5^s\}$ y en azul los datos $\mathbf{X}^t = \{\mathbf{x}_1^t, \mathbf{x}_2^t\}$. Las medidas μ_s y μ_t son uniformes sobre estos datos. Las flechas indican el transporte óptimo γ_0 obtenido al utilizar el cuadrado de la distancia euclídea como función de costo. La medida de los datos transportados es $\hat{\mu} = \frac{1}{2}\delta_{\mathbf{x}_1^t} + \frac{1}{2}\delta_{\mathbf{x}_2^t}$.

viene dado por la distancia al cuadrado entre los puntos, específicamente el plan de transporte óptimo es

$$\gamma_0 = \begin{pmatrix} \frac{1}{6} & 0 \\ \frac{1}{6} & 0 \\ \frac{1}{6} & 0 \\ 0 & \frac{1}{4} \\ 0 & \frac{1}{4} \end{pmatrix}$$

La figura 4.3 muestra la ubicación de los conjuntos $\mathbf{X}^s$ y $\mathbf{X}^t$ así como el transporte óptimo γ_0. A partir del transporte óptimo tenemos que

$$\hat{p}_1 = \sum_{i=1}^{5} \gamma_0(i,1) = \frac{1}{2} \quad \hat{p}_2 = \sum_{i=1}^{5} \gamma_0(i,2) = \frac{1}{2},$$

por lo tanto los datos transportados conforman la medida

$$\hat{\mu} = \frac{1}{2}\delta_{(2,2)} + \frac{1}{2}\delta_{(5,2)}.$$

A partir de lo anterior podemos calcular una transformación de los datos $\mathbf{X}^s$, lo cual lleva a la siguiente definición.

Definición 4.8 (Barycentric Mapping). *Dado un dato $\boldsymbol{x}_i^s \in \boldsymbol{X}^s$, su mapeo*

baricéntrico es

$$\hat{\boldsymbol{x}_i^s} := \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \sum_j \gamma_0(i,j)c(\boldsymbol{x}, \boldsymbol{x}_j^t).$$

La elección del dato $\mathbf{x}_i^s$ fija el valor de i y por lo tanto la fila de $\gamma_0(i,\cdot)$. Luego, se busca el punto $\mathbf{x} \in \mathbb{R}^n$ que minimiza la cantidad $\sum_j \gamma_0(i,j)c(\mathbf{x}, \mathbf{x}_j^t)$, con i fijo. El dato $\hat{\mathbf{x}}_i^s$ es el punto a donde hay que transportar el dato $\mathbf{x}_i^s$ para minimizar el costo. Al utilizar el cuadrado de la distancia euclídea como función de costo, el mapeo baricéntrico para todo el conjunto $\mathbf{X}^s$ puede expresarse como

$$\hat{\mathbf{X}}^s = T_{\gamma_0}(\mathbf{X}^s) := diag(\gamma_0 \mathbf{1}_{n_t})^{-1}\gamma_0 \mathbf{X}^t,$$

donde, abusando de la notación, $\mathbf{X}^t$ es un vector conformada por los datos (en lugar de un conjunto), explícitamente

$$\mathbf{X}^t = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_{n_t}^t \end{pmatrix},$$

siendo n_t el cardinal de $\mathbf{X}^t$ (como conjunto).

Retomando el ejemplo 4.7, tenemos que $n_t = 2$ y

$$\gamma_0 \mathbf{1}_{n_t} = \begin{pmatrix} 1/6 & 0 \\ 1/6 & 0 \\ 1/6 & 0 \\ 0 & 1/4 \\ 0 & 1/4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/4 \\ 1/4 \end{pmatrix}.$$

Luego como la inversa de una matriz diagonal con entradas no nulas $diag(a_1, \ldots, a_n)$ es otra matriz diagonal de la forma $diag(\frac{1}{a_1}, \ldots, \frac{1}{a_n})$ tenemos que

$$diag(\gamma_0 \mathbf{1}_{n_t})^{-1} = \begin{pmatrix} 6 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Por lo tanto,

$$diag(\gamma_0 \mathbf{1}_{n_t})^{-1}\gamma_0 = \begin{pmatrix} 6 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 1/6 & 0 \\ 1/6 & 0 \\ 1/6 & 0 \\ 0 & 1/4 \\ 0 & 1/4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Finalmente, tenemos que el mapeo baricéntrico es

$$\hat{\mathbf{X}}^s = diag(\gamma_0 \mathbf{1}_{n_t})^{-1}\gamma_0 \mathbf{X}^t = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 5 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \\ 5 & 2 \\ 5 & 2 \end{pmatrix}.$$

Observamos que reencontramos el resultado obtenido en el ejemplo 4.7, ya que lo anterior nos dice que los primeros tres puntos del conjunto source $\mathbf{x}_1^s, \mathbf{x}_2^s, \mathbf{x}_3^s$ se transportan a $(2,2) = \mathbf{x}_1^t$ mientras que los últimos dos puntos del conjunto source $\mathbf{x}_4^s$ y $\mathbf{x}_5^s$ se transportan a $(5,2) = \mathbf{x}_2^t$. En conclusión, el mapping baricéntrico nos da una manera de transportar los datos del conjunto source hasta el conjunto target utilizando el transporte óptimo y operaciones matriciales.

Un resultado importante presentado en el artículo "Fast Computation of Wasserstein Barycenters" de Cuturi y Doucet (2014), es que cuando las medidas μ_s y μ_t son uniformes (posiblemente con distinta cantidad de puntos) entonces el mapping baricéntrico se reduce a

$$\hat{\mathbf{X}}^s = n_s \gamma_0 \mathbf{X}^t \quad \text{y} \quad \hat{\mathbf{X}}^t = n_t \gamma_0^T \mathbf{X}^s.$$

Este resultado será sumamente útil en los experimentos, dado que bajo la falta de información externa, siempre supondremos que cada dato de la muestra tiene el mismo peso, considerando así que provienen de distribuciones uniformes.

Un aporte interesante de Courty, Flamary, Tuia y Rakotomamonjy (2016) es el de agregar un factor de regularización para preservar la información de las etiquetas. Al utilizar el transporte óptimo tanto en la formulación de Monge como la de Kantorovich, o en la versión regularizada, suele suceder que los

datos transportados queden del lado incorrecto con respecto a sus etiquetas. Por ejemplo, si en la figura 4.3 suponemos que todos los datos que están a la derecha de la recta $x = 2.5$ son de una clase y los que están a la izquierda son de otra, tenemos que el punto $(3, 3)$ cambia de clase. Un ejemplo real de este fenómeno puede verse en el experimento sobre MNIST, en el capítulo 5. Ellos proponen penalizar los transportes que transportan datos de $\mathbf{X}^s$ con diferentes etiquetas al mismo dato de $\mathbf{X}^t$. Para esto introducen una nueva regularización:

Definición 4.9 (Regularización con etiquetas de clase). *El problema de transporte óptimo con regularización de clase es*

$$\gamma_0 = \underset{\gamma \in \Pi(\mu,\nu)}{\arg\min} \underbrace{\langle \gamma, \boldsymbol{C} \rangle_F + \epsilon H(\gamma)}_{\textit{Transporte óptimo regularizado}} + \underbrace{\eta \mathcal{R}_c(\gamma)}_{\textit{Regularización de clase}}$$

.

Proponen dos posibles regularizaciones de clase: Regularización "group-sparsity" y Regularización Laplaciana.

La intuición detrás del método regularización Group-sparsity es que queremos que cada dato de target $\mathbf{x}_i^t$ reciba masa únicamente de datos de source $\mathbf{x}_i^s$ que pertenezcan a la misma clase. La función de regularización en este caso es

$$\mathcal{R}_c(\gamma) = \sum_j \sum_{cl} ||\gamma(\mathcal{I}_{cl}, j)||_2,$$

donde $\mathcal{I}_{cl}$ es el conjunto de índices de las filas de γ correspondientes con datos de *source* $\mathbf{x}_i^s$ de la clase cl. Luego $\gamma(\mathcal{I}_{cl}, j)$ es un vector que contiene los coeficientes de la j-ésima columna de γ asociados a la clase cl. Una hipótesis importante para que este tipo de regularización tenga buenos resultados es que las distribuciones de las etiquetas sean muy similares, es decir que $\mathbf{P}_s(y)$ y $\mathbf{P}_t(y)$ sean funciones cercanas.

En la regularización Laplaciana apuntamos a preservar la estructura de los datos (aproximada por un grafo) durante el transporte. Las ideas detrás de este tipo de regularización se presentaron en el artículo "Regularized Discrete Optimal Transport" de Ferradans, Papadakis, Rabin, Peyré y Aujol (2013). La intuición detrás es que es deseable que datos cercanos, en términos de la distancia, en el conjunto source $\mathbf{X}^s$ se transporten a datos cercanos en el conjunto target $\mathbf{X}^t$. Si $\mathbf{x}_i^s$ es un dato del source, denotaremos como antes, $\hat{\mathbf{x}}_i^s$ al

dato transportado. Dada una matriz de semejanza no negativa $\mathbf{S} = (s(i,j))_{i,j}$ sobre los datos del conjunto source $\mathbf{X}^s$, definimos la función de regularización como

$$\mathcal{R}_c(\gamma) = \frac{1}{n_s^2} \sum_{i,j} s(i,j) ||\hat{\mathbf{x}}_i^s - \hat{\mathbf{x}}_j^s||_2^2.$$

En la práctica, podemos definir $s(i,j) = 0$ si $y_i^s \neq y_j^s$.

Cuando las medidas μ_s y μ_t sean ambas uniformes, la función de regularización Laplaciana se simplifica a

$$\mathcal{R}_c(\gamma) = Tr((\mathbf{X}^t)^T \gamma^T \mathbf{L} \gamma \mathbf{X}^t),$$

donde $\mathbf{L} = diag(\mathbf{S} - \mathbf{1}) - \mathbf{S}$.

Capítulo 5

Experimentos

Los experimentos juegan un papel crucial en la validación y demostración de la eficacia de los métodos estudiados sobre transporte óptimo y la adaptación de dominio. A través de experimentos, es posible evaluar el rendimiento de los algoritmos en diversos escenarios y bajo diferentes condiciones de datos. Este capítulo presenta una serie de experimentos realizados en Python que ilustran la aplicación práctica de los conceptos teóricos discutidos previamente en esta tesis.

En particular, se llevarán a cabo experimentos en tres áreas principales: regresión lineal, clasificación de dígitos y transferencia de color. Cada una de estas áreas representa un tipo diferente de desafío en la adaptación de dominio, permitiendo así una evaluación exhaustiva de las capacidades y limitaciones de los métodos basados en transporte óptimo.

- **Regresión Lineal:** buscaremos verificar que tan útil puede ser el transporte óptimo cuando necesitamos rotar una regresión lineal. Este es un ejemplo de adaptación de dominio supervisado (SDA).
- **Clasificación de Dígitos:** en este experimento, utilizaremos conjuntos de datos de dígitos escritos a mano para investigar cómo el transporte óptimo puede mejorar la precisión de los modelos de clasificación cuando se enfrentan a variaciones estilísticas entre diferentes conjuntos de datos. Este es un ejemplo de adaptación de dominio no supervisado (UDA).
- **Transferencia de Color:** exploraremos cómo los métodos de transporte óptimo pueden ser utilizados para la transferencia de estilos de color entre imágenes, un problema relevante en aplicaciones de procesamiento de imágenes y visión por computadora.

A lo largo de este capítulo, se describirán en detalle los procedimientos experimentales, se presentarán los resultados obtenidos y se discutirán las implicancias de estos hallazgos. La implementación y los parámetros específicos utilizados en cada simulación se documentarán cuidadosamente para garantizar la reproducibilidad de los experimentos.

El objetivo de estos experimentos es no solo validar los métodos teóricos propuestos, sino también proporcionar una guía práctica sobre cómo aplicar técnicas de transporte óptimo en problemas reales de adaptación de dominio. Los resultados obtenidos demuestran el potencial y la versatilidad del transporte óptimo como herramienta para la adaptación de dominio, destacando su aplicabilidad en una amplia gama de escenarios. El código de todos los experimentos lo podrán encontrar en GitHub.

5.1. Regresión lineal

En esta sección mostraremos un resultado obtenido en el transcurso de esta tesis, en donde utilizando la geometría del transporte óptimo buscamos adaptar una regresión lineal cuando el dominio target $\mathcal{X}_t$ consiste en una rotación del dominio source $\mathcal{X}_s$. Específicamente buscamos estudiar la siguiente situación: contamos con una gran cantidad de observaciones en el dominio source las cuales usamos para ajustar una regresión lineal. Luego, las observaciones del dominio target se desvían mediante una rotación R_θ. Además, supondremos que la cantidad de estas observaciones es mucho menor que las observaciones iniciales. Para simplificar las simulaciones, suponemos que ambas muestras se obtienen de manera lineal a partir de $Y = aX + \epsilon$ con $\epsilon \sim \mathcal{N}(0, \sigma^2)$ ruido gaussiano.

Simularemos n_s observaciones del conjunto source $\mathbf{X}_s \times Y_s = \{(x_i^s, y_i^s) | x_i^s, y_i^s \in \mathbb{R}\}_{i=1}^{n_s} \subset \mathbb{R} \times \mathbb{R}$ y n_t observaciones del conjunto target $\mathbf{X}_t \times Y_t = \{(x_j^t, y_j^t) | x_j^t, y_j^t \in \mathbb{R}\}_{j=1}^{n_t} \subset \mathbb{R} \times \mathbb{R}$, donde $n_t \ll n_s$. Como en la mayoría de las aplicaciones, utilizaremos la discretización Lagrangiana (ver sección 3.3), por lo tanto μ y ν son dos medidas discretas uniformes:

$$\mu = \sum_{i=1}^{n_s} \frac{1}{n_s} \delta_{(x_i^s, y_i^s)} \quad \text{y} \quad \nu = \sum_{j=1}^{n_t} \frac{1}{n_t} \delta_{(x_j^t, y_j^t)}.$$

La figura 5.1 muestra esta situación donde fijamos $n_s = 500$, $n_t = 10$, el

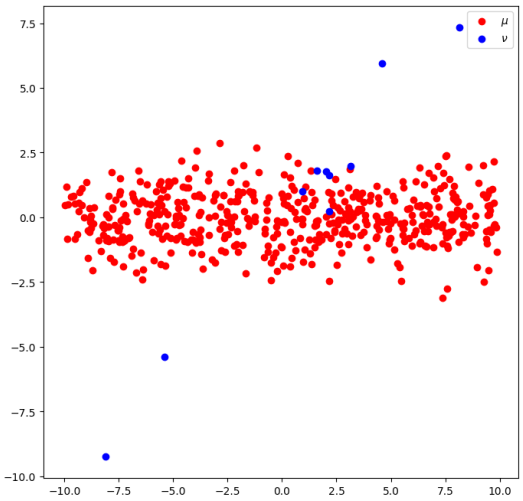


Figura 5.1: Simulaciones de las observaciones descritas en la sección 5.1. En este caso $n_s = 500$, $n_t = 10$, el ruido gaussiano de ambas muestras tiene desvió $\sigma = 1$ y el ángulo de rotación es $\pi/4$.

desvió del ruido gaussiano es $\sigma = 1$ en ambos casos. La pendiente del modelo donde se sortean las muestras del conjunto source es $a_s = 0$ y el ángulo de rotación es $\theta = \pi/4$, y por lo tanto la pendiente del modelo donde se sortean las muestras del conjunto target es $a_t = 1$.

Es sabido que la estimación obtenida a partir de una regresión lineal en general se favorece al contar con mayor cantidad de datos, es decir, tener más datos suele implicar obtener una mejor estimación de los parámetros. Con esto en mente, queremos utilizar la información en source y lo que sabemos en cuanto a como se relacionan el conjunto target con el original, para adaptar la regresión lineal ajustada con los datos del conjunto source.

Estudiaremos este problema con dos variantes:

1. Suponiendo que conocemos el ángulo real θ que relaciona ambos dominios.
2. Estimando el ángulo θ utilizando el transporte óptimo.

Variante 1 (ángulo conocido)

Para simplificar las cuentas supondremos que $n_s = n_t$ y no hay ruido, lo cual es equivalente a que $\sigma = 0$. Luego mostraremos que estas suposiciones no son necesarias. La figura 5.2 muestra una posible situación.

Conociendo los parámetros a_s y b_s del modelo lineal de donde se sortean

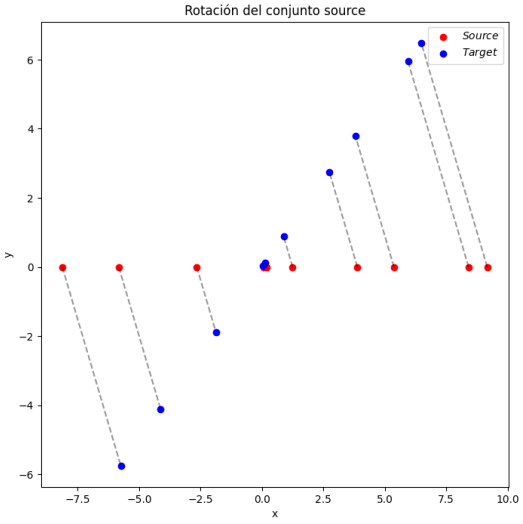


Figura 5.2: El conjunto target se obtiene del source mediante un rotación de ángulo $\theta = \pi/4$. Las lineas punteadas muestran como se transforma cada punto.

los datos del conjunto source, nos proponemos estimar los parámetros a_t y b_t para que la recta $\hat{y} = a_t\hat{x} + b_t$ ajuste los datos del conjunto target, donde $(\hat{x}, \hat{y})^T = R_\theta(x, y)^T$ con R_θ la matriz de rotación de ángulo θ. Por lo tanto, tenemos que $\hat{x} = xcos\theta - ysen\theta$ y $\hat{y} = xsen\theta + ycos\theta$. Reemplazando lo anterior en $\hat{y} = a_t\hat{x} + b_t$ obtenemos

$$xsen\theta + ysen\theta = a_t\Big(xcos\theta - ysen\theta\Big) + b_t$$

$$y\Big(cos\theta + a_tsen\theta\Big) = \Big(a_tcos\theta - sen\theta\Big)x + b_t$$

$$y = \underbrace{\frac{a_tcos\theta - sen\theta}{cos\theta + a_tsen\theta}}_{a_s} x + \underbrace{\frac{b_t}{cos\theta + a_tsen\theta}}_{b_s}$$

Despejando a_t y b_t de la ecuación anterior obtenemos

$$a_t = \frac{a_scos\theta + sen\theta}{cos\theta - a_ssen\theta} \tag{5.1}$$

$$b_t = b_s\Big(cos\theta + a_tsen\theta\Big) \tag{5.2}$$

Como conocemos a_s, b_s y θ podemos calcular los coeficientes de la regresión

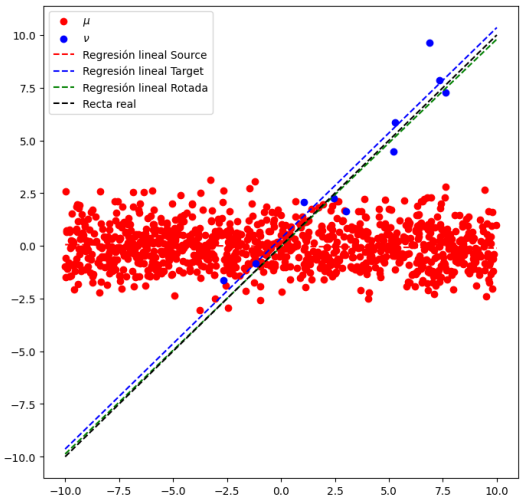


Figura 5.3: Resultado del procedimiento propuesto cuando se conoce el ángulo θ. En azul la recta de regresión target, en negro la verdadera y en verde la obtenida mediante el procedimiento propuesto. En esta simulación se obtuvo un error cuadrático medio (MSE) de 0.128 y 0.010 para la recta azul y verde respectivamente.

rotada que mejor ajusta a las observaciones del conjunto target. Volviendo al escenario donde $n_s \gg n_t$ proponemos el siguiente método para mejorar la estimación en conjunto target:

1. Utilizar los datos $\{x_i^s, y_i^s\}_{i=1}^{n_s}$ para estimar los coeficientes a_s y b_s utilizando el algoritmo de regresión lineal.
2. A partir de a_s, b_s y θ calcular los nuevos coeficientes a_t y b_t.
3. Utilizar la recta $y = a_t x + b_t$ para ajustar los datos $\{x_j^t, y_j^t\}_{j=1}^{n_t}$.

La figura 5.3 muestra el resultado de este procedimiento para una simulación. Para obtener información más consistente realizamos el siguiente experimento:

- Fijamos $n_t = 10$, $\sigma = 1$ y $\theta = \pi/4$.
- Variamos n_s en $\{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$.
- Para cada valor de n_s realizamos 1000 sorteos de los conjuntos source y target y realizamos el procedimiento.
- Calculamos el error cuadrático medio (MSE) promedio para cada n_s así como su varianza.

La figura 5.4 muestra los resultados obtenidos. Vemos que cuando la cantidad de datos en el conjunto source es similar a la cantidad de datos en el conjunto target entonces este procedimiento no da mejores resultados. Por otro

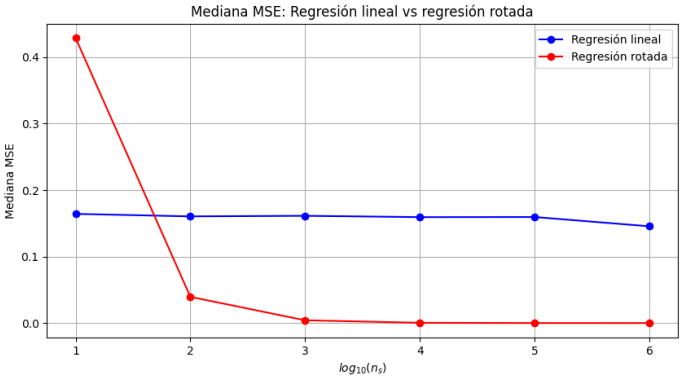


Figura 5.4: Resultado del proceso de simular 1000 veces el procedimiento propuesto para distintos valores de n_s. El eje x está en escala logarítmica. El procedimiento reporta un MSE promedio menor en la situación donde $n_s \gg n_t$.

lado, al aumentar la cantidad de datos en source apreciamos una mejora en la estimación de la recta de ajuste. Proporcionamos además los valores numéricos en la tabla 5.1 para comprobar que el error cuadrático medio disminuye al aumentar n_s.

Mediana MSE		
$log_{10}(n_s)$	**Regresión lineal**	**Regresión rotada**
1	0.164187	0.428760
2	0.160415	0.039448
3	0.161311	0.004191
4	0.159287	0.000435
5	0.159465	0.000038
6	0.145519	0.000004

Tabla 5.1: Mediana MSE: Regresión lineal vs regresión rotada

Si bien es claro que en promedio el procedimiento mejora la estimación es de interés tener información sobre que tanto puede variar el resultado del procedimiento de una muestra a otra, para esto realizamos un boxplot de la simulación anterior para tener información de la varianza del error cuadrático medio (MSE). La figura 5.5 muestra dicho boxplot. A medida que n_s aumenta, no solo el MSE promedio obtenido es inferior al que obtenemos estimando la regresión utilizando los datos del conjunto target, si no que además, la varianza de dicha estimación también decrece. Concluimos que si conocemos el ángulo de rotación que relaciona ambos conjuntos entonces el procedimiento propuesto es sumamente útil.

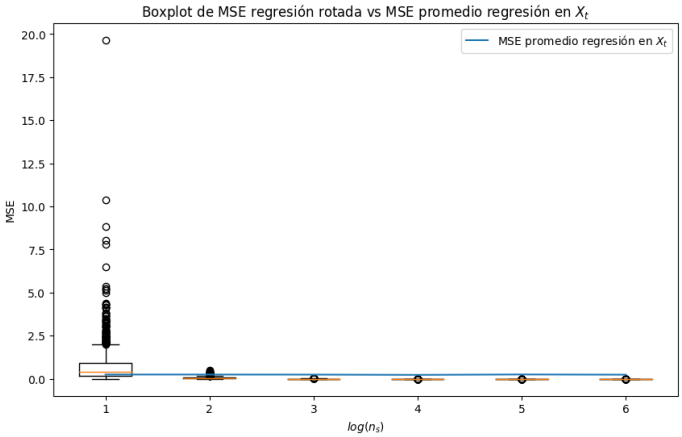


Figura 5.5: Boxplot del proceso de simular 1000 veces el procedimiento propuesto para distintos valores de n_s. El eje x está en escala logarítmica. El procedimiento propuesto no solo proporciona una mejor estimación en promedio, si no que además, la variación de las estimaciones decrece al aumentar la cantidad de datos en source n_s.

Variante 2 (ángulo estimado)

Si bien en la sección anterior comprobamos que el método propuesto es robusto y da mejores resultados que simplemente estimar los parámetros de una regresión con los datos del conjunto target, fue fundamental conocer el ángulo de rotación θ que vincula los conjuntos source y target. En las aplicaciones reales rara vez conocemos el ángulo de antemano, por esta razón proponemos una solución utilizando transporte óptimo para estimar el ángulo θ y luego utilizarlo con el procedimiento descrito en la sección anterior.

Probamos en la proposición 2.15 que si el conjunto target se relaciona con el conjunto source mediante una rotación entonces la asignación que se obtiene del transporte óptimo coincide con la rotación. Aún en ausencia de ruido, en la situación donde $n_s \gg n_t$ tenemos que el conjunto target no es exactamente el conjunto source rotado. Para solucionar este problema, primero utilizamos $K-$means en el conjunto source eligiendo la cantidad de grupos como $K = n_t$ para así obtener un subconjunto, los centroides, $\{(\hat{x}_1^s, \hat{y}_1^s), \ldots, (\hat{x}_{n_t}^s, \hat{y}_{n_t}^s)\}$ que es representativo del conjunto source pero que tiene la misma cantidad de puntos que el conjunto target. Al conjunto obtenido con este procedimiento lo llamaremos conjunto K-source y lo denotaremos X_{K_s}.

Luego del paso anterior, tenemos dos conjuntos $\{(\hat{x}_1^s, \hat{y}_1^s), \ldots, (\hat{x}_{n_t}^s, \hat{y}_{n_t}^s)\}$ y $\{(x_1^t, y_1^t), \ldots, (x_{n_t}^t, y_{n_t}^t)\}$ con la misma cantidad de puntos. El siguiente paso

es calcular el transporte óptimo entre el conjunto K-source y el conjunto target, utilizando como función de costo la distancia euclídea. Para simplificar la notación supondremos (a lo sumo cambiando los índices) que los conjuntos K-source y target están ordenados según el transporte óptimo T, es decir, $T((\hat{x}_i^s, \hat{y}_i^s)) = (x_i^t, y_i^t)$ para todo $i = 1, \ldots, n$.

Luego de conocer a cual punto del conjunto target se mapea cada punto del conjunto K-source, como sabemos que el transporte óptimo coincide con la rotación, podemos estimar el ángulo de rotación utilizando la descomposición en valores singulares como describimos al final del capítulo 1. Abusando de la notación, podemos suponer que X_{K_s} y $T(X_{K_s})$ son representaciones matriciales del conjunto K-source y su transportado. A partir de esto, la matriz de similitud en este contexto es

$$H = X_{K_s}^T T(X_{K_s}),$$

la cual representa la similitud entre los centroides obtenidos mediante K-means y sus transportados. Luego, calculando la descomposición en valores singulares podemos escribir $H = U\Sigma V^T$ y por lo tanto podemos estimar la matriz de rotación como $\hat{R}_\theta = VU^T \in \mathcal{M}_{2\times 2}$.

La primera columna de $\hat{R}$ es una aproximación del vector $z = (cos\theta, sen\theta)$ por lo que podemos aproximar el ángulo θ como el argumento del complejo z. En las simulaciones, este último paso lo hacemos utilizando la función *atan* de *numpy*.

Usaremos la estimación $\hat{\theta}$ para luego aplicar la variante 1. La figura 5.6 muestra los resultados obtenidos al realizar la misma simulación que en la variante 1 pero estimando el ángulo mediante el procedimiento anterior. Observamos que aumentar la cantidad de observaciones en el conjunto source da mejores estimaciones en mediana aunque no es una mejora tan significativa como antes, esto era esperable, dado que al estimar el ángulo estamos introduciendo más error en el modelo.

Al igual que antes, proporcionamos los valores números obtenidos en la simulación en la figura 5.2. Observamos que es necesario una mayor cantidad de datos de source n_s para obtener mejoras relevantes.

El pseudo código 3 compila el procedimiento propuesto para estimar el ángulo de rotación entre los conjuntos source y target.

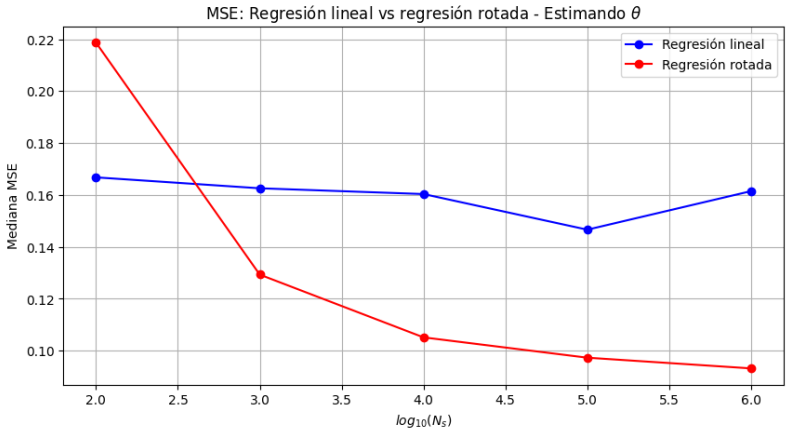


Figura 5.6: Resultados obtenidos al realizar la simulación de la variante 1 pero estimando el ángulo θ. Al aumentar la cantidad de observaciones en el conjunto source se ve una mejora en la estimación.

Algorithm 3 Estimación del ángulo de rotación utilizando K-means, transporte óptimo y SVD

Require: Coordenadas X_s, X_t, número de centroides nt
Ensure: Estimación del ángulo de rotación θ

Paso 1: Aplicar k-means en X_s
$centroides \leftarrow \text{KMeans}(X_s, \text{n\_clusters} = nt)$

Paso 2: Resolver el problema de Monge
$C \leftarrow$ Matriz de costo con la distancia euclídea
$P_\gamma \leftarrow$ Resolver el problema de Monge usando medidas μ y ν uniformes.

Paso 3: Mapear los centroides de X_s a X_t
$centroides\_transportados \leftarrow$ Transportar los centroides usando P_γ

Paso 4: Usar SVD para estimar la matriz de rotación
$\hat{R} \leftarrow$ Aplicar el algoritmo con SVD para estimar la matriz de rotación usando SVD

Paso 5: Estimar el ángulo de rotación
$\theta \leftarrow \arctan2(\hat{R}[1,0], \hat{R}[0,0])$
return θ

Mediana MSE - Estimación del ángulo		
$log_{10}(n_s)$	**Regresión lineal**	**Regresión rotada**
2	0.166780	0.218713
3	0.162574	0.129299
4	0.160353	0.105116
5	0.146619	0.097287
6	0.161478	0.093161

Tabla 5.2: Mediana MSE al estimar el ángulo: Regresión lineal vs regresión rotada

Situaciones reales

A pesar de que las simulaciones muestran un resultado alentador, en situaciones reales no podemos simular miles de datos, si no que el conjunto source y target están fijos. Para adaptar el procedimiento anterior a este contexto proponemos el siguiente camino:

1. Dejar fijo el conjunto target.
2. Elegir un número p entre $(0.5, 0.8)$ que usaremos como proporción.
3. Sortear $p \times n_s$ observaciones sin reposición del conjunto source.
4. Utilizando el conjunto target y el obtenido en el paso anterior como conjunto source para aplicar el procedimiento descrito en la sección anterior.
5. Repetir los pasos 3 y 4 una cantidad N de veces.
6. Promediar los parámetros a_r y b_r obtenidos en las N repeticiones.

Para comprobar este procedimiento realizamos un experimento con los siguientes parámetros:

- $n_s = 1000$ y $n_t = 10$
- $\sigma = 1$
- $a_s = 0$ (es decir, los datos de source siguen la recta $y = 0 + \epsilon$).
- $p = 0.8$
- $N = 2000$

Elegimos cuatro ángulos para simular: $\theta \in \{\pi/6, \pi/5, \pi/4, \pi/3\}$. Para cada uno de estos ángulos realizamos 50 realizaciones del proceso con los parámetros especificados anteriormente. La figura 5.7 muestra el error cuadrático medio (MSE) de la regresión calculada en el conjunto target X_t y de la regresión

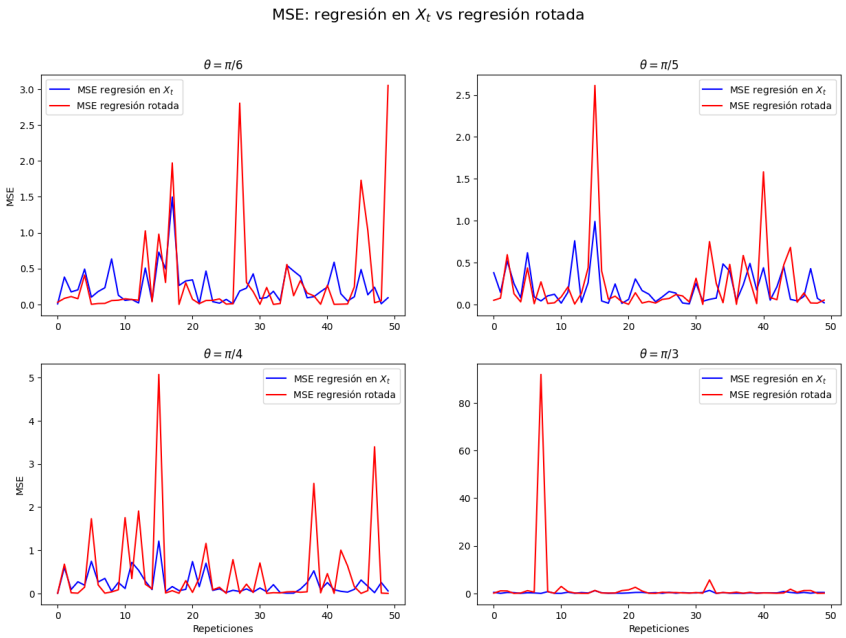


Figura 5.7: Resultados del procedimiento propuesto para diferentes valores de θ. El procedimiento da mejores resultados en algunas situaciones pero tiene grandes outliers (picos altos en rojo).

rotada. El método propuesto da mejores resultados si la curva roja está por debajo de la curva azul.

Observamos en la figura 5.7 que el procedimiento propuesto puede dar mejores resultados pero también corremos el riesgo de obtener resultados muchos peores, como muestran los picos en rojos. Para entender un poco más esta variabilidad decidimos realizar gráficos de caja. El resultado se muestra en la figura 5.8. Concluimos que para ángulos pequeños como $\pi/6$ y $\pi/5$ el procedimiento parece ser estable y da en promedio mejores resultados. Por otro lado, para ángulos grandes como $\pi/3$ el procedimiento puede dar mejores resultados, pero sufre de una gran variación, reduciendo el uso del mismo en aplicaciones reales. El pseudo-código 4 representa el procedimiento planteado.

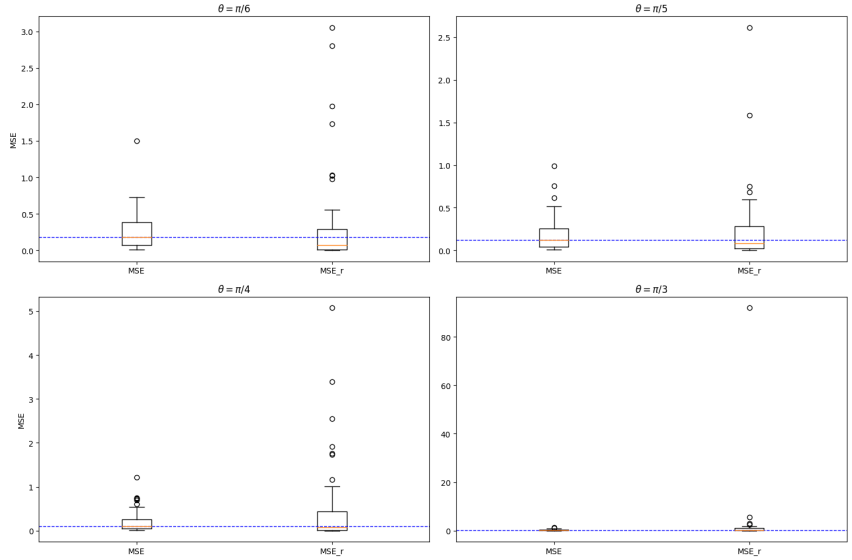


Figura 5.8: Gráficos de caja obtenidos para la simulaciones mostradas en la figura 5.7. La recta azul está a la altura de la mediana de los errores cuadráticos medios obtenidos al calcular la regresión con los datos del conjunto target X_t. Observamos que para ángulos pequeños como $\pi/6$ y $\pi/5$ da mejores resultados. Para ángulos mayores como $\pi/3$ el procedimiento tiene mayor varianza, reduciendo su uso en aplicaciones reales.

Algorithm 4 Estimación de los parámetros de la regresión rotada

Require: X_s, X_t, número de repeticiones n_{rep}, proporción p
Ensure: a_r, b_r
  Inicializar listas A_r, B_r
  $ns \leftarrow \text{len}(X_s)$
  $nt \leftarrow \text{len}(X_t)$
  for *sorteo* in range(n_{rep}) **do**
    $X_{Ks} \leftarrow$ Aplicar bootstrap sobre X_s con proporción p
    $\theta \leftarrow$ aplicar algoritmo 3 para estimar el ángulo
    $a_1, b_1 \leftarrow \text{regresion\_lineal}(x1, y1)$
    $a_2, b_2 \leftarrow \text{regresion\_rotada}(a_1, b_1, \theta)$
    $A_r \leftarrow \text{append}(a_2)$
    $B_r \leftarrow \text{append}(b_2)$
  end for
  $a_r \leftarrow$ promedio de A_r
  $b_r \leftarrow$ promedio de B_r
  return a_r, b_r

5.2. Clasificación de dígitos

En esta sección veremos como aplicar la adaptación de dominio no supervisada en un problema de clasificación. Un conjunto de datos muy utilizado cuando se comienza a aprender sobre machine learning es el conocido **MNIST**, que cuenta con 70.000 imágenes de los dígitos $0, 1, 2, 3, 4, 5, 6, 7, 8$ y 9 escritos a mano. Las mismas tienen un tamaño de 28×28 pixeles y están en escala de grises. Otro conjunto de datos similar, pero no tan conocido, es el **USPS** (Hull, 1994), el cual consiste en 7900 imágenes de dígitos, también escritos a mano y en escala de grises pero de tamaño 16×16. Entrenaremos como clasificador un support vector machine (SVM) para la tarea de clasificar el conjunto de datos **MNIST**. Veremos que el modelo encontrado no tiene buenos resultados cuando lo aplicamos en el dataset **USPS** pero que si aplicamos transporte óptimo podemos mejorar el resultado.

Como preprocesamiento se realiza un zero-padding (agregar 0 en los bordes) en las imágenes de **USPS** para que tengan el mismo tamaño que **MNIST**, es decir, 28×28 píxeles. En la figura 5.9 mostramos algunas de las observaciones, vemos claramente que algunos dígitos tienen mayores diferencias entre ambos conjuntos, por ejemplo los dígitos $1, 2$ y 5. Cada imagen está representado por un punto en $\mathbb{R}^{28\times 28} = \mathbb{R}^{784}$.

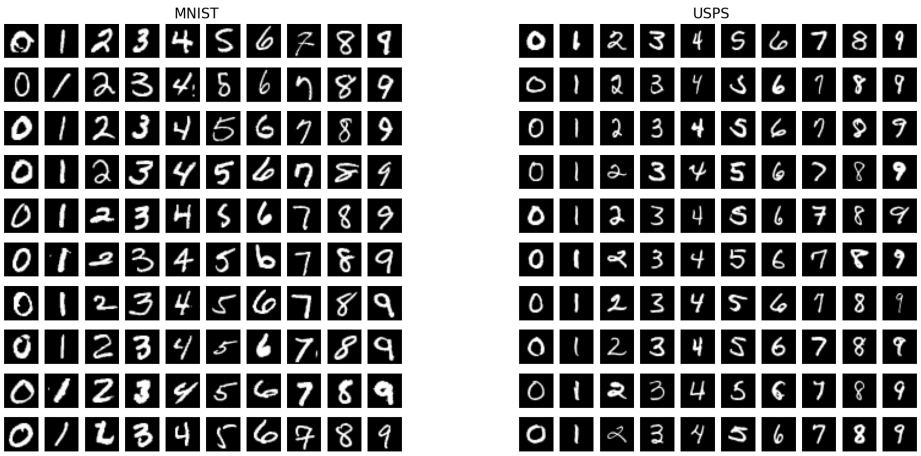


Figura 5.9: Ejemplos de los dataset MNIST (izquierda) y USPS (derecha) de reconocimiento de digitos

Para ganar un poco de intuición sobre cómo se agrupan los dígitos en ambos conjuntos de datos, podemos utilizar la técnica de reducción de dimensionalidad conocida como t-SNE (t-distributed Stochastic Neighbor Embedding -

van der Maaten y Hinton (2008)) para visualizar los datos en el plano. La figura 5.10 muestra el resultado de aplicar t-SNE a ambos conjuntos de datos: **MNIST** y **USPS**. Podemos observar que cada distribución de puntos presenta sus clases relativamente bien agrupadas. Esto sugiere que las características intrínsecas de los datos permiten una clasificación clara y precisa dentro de cada grupo.

Sin embargo, es evidente que las clases de cada uno de los conjuntos están separadas entre si. Esta separación intergrupal sugiere que las diferencias entre las clases son lo suficientemente pronunciadas como para evitar solapamientos significativos entre ellas. Por ejemplo, si miramos la clase del dígito 1 (en naranja en la figura 5.10) vemos que no tienen puntos en común, así un clasificador que logre clasificar el dígito 1 del conjunto de datos **MNIST** no tiene por qué clasificar bien el mismo dígito en el conjunto de dato **USPS**.

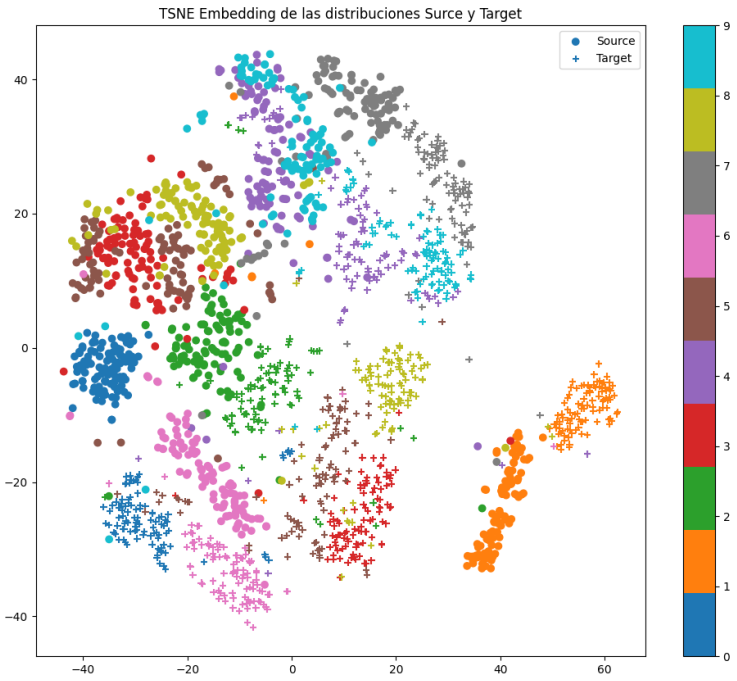


Figura 5.10: Proyección en el plano mediante la técnica TSNE de MNIST y USPS.

A continuación procedemos a calcular la matriz de costo entre las observaciones del conjunto source y el conjunto target, para la cual utilizamos la distancia euclídea como función de costo. La figura 5.11 muestra la matriz de costo, observamos por ejemplo que el dígito 1 del dataset **USPS** es sustancialmente diferente al resto de dígitos, esto podemos constatarlo con una

inspección visual sobre la figura 5.9. Para calcular el transporte óptimo necesitamos trabajar con distribuciones de probabilidad, por lo que utilizaremos la discretización Lagrangiana, es decir, tenemos una medida uniforme en $\mathbb{R}^{784}$ para cada conjunto.

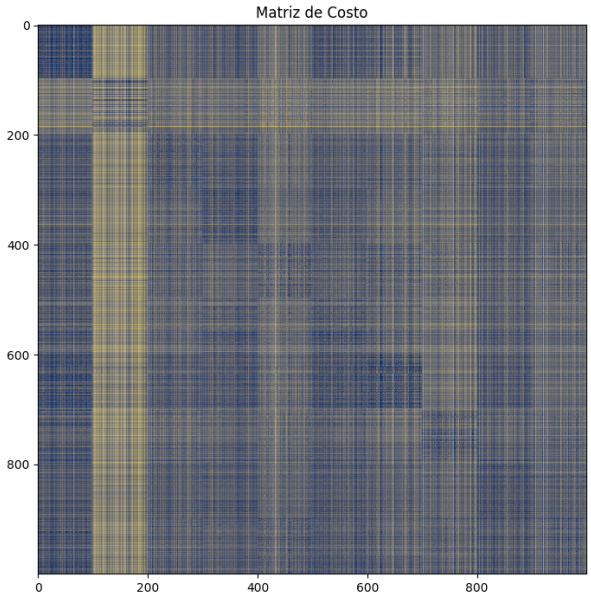


Figura 5.11: Matriz de costo entre **MNIST** y **USPS** utilizando la distancia Euclídea.

La figura 5.12 muestra los planes de transporte obtenidos para el problema de Kantorovich y el problema regularizado. Observaremos que la mayoría de los transportes no nulos se concentran en la diagonal, esta disposición indica que, en promedio, cada clase tiende a transportarse a su clase correspondiente de manera bastante precisa. Además es evidente que el plan de transporte del problema regularizado es más disperso que la versión sin regularizar, situación esperable por lo detallada en la sección 2.4.

A continuación entrenamos un clasificador basado en Support Vector Machine en el dataset **MNIST** y el dataset **MNIST** transportado utilizando el plan de transporte, la tabla 5.3 muestra la tasa de acierto obtenido:

Hasta el momento solo utilizamos el transporte óptimo proveniente del problema de Kantorovich, pero surge la pregunta de que tanto se podrá mejorar utilizando la versión regularizada. La figura 5.13 muestra lo que sucede al aumentar el factor de regularización. "Accuracy naive" hace referencia a no aplicar ninguna técnica de transporte óptimo.

Para entender un poco mejor por qué el aplicar transporte óptimo en

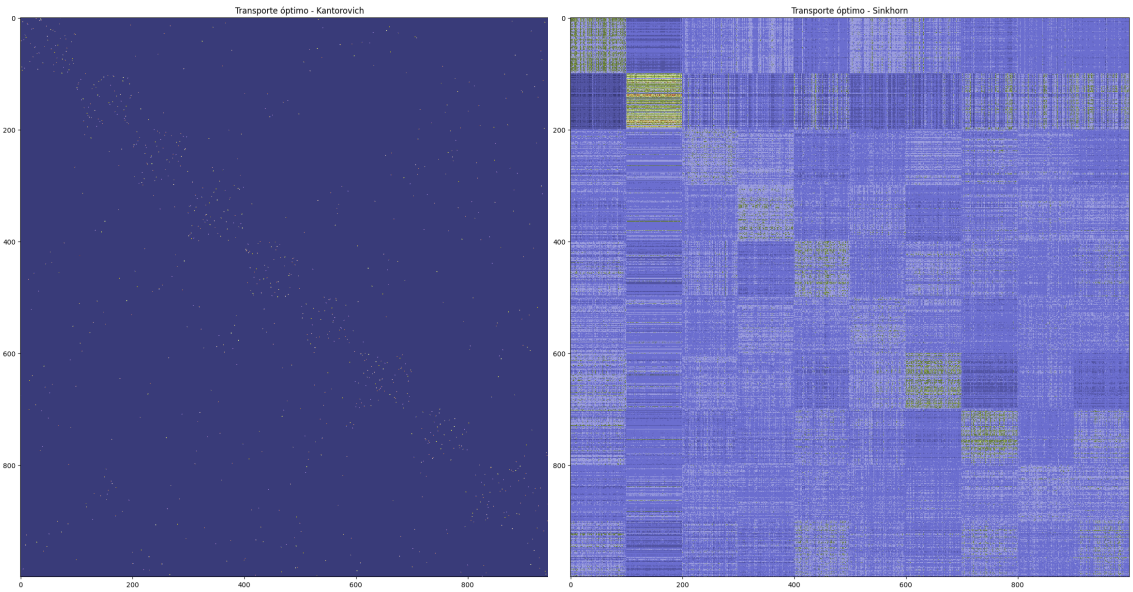


Figura 5.12: Planes de transporte obtenidos. A la izquierda el de Kantorovich y a la derecha el de Sinkhorn. Los colores azules representa menores valores de γ_{xy} mientras que los tonos amarillos representan valores más altos.

Conjunto	Tasa de acierto
MNIST	0.978
USPS	0.687
USPS (con transporte)	0.764

Tabla 5.3: Tasa de acierto obtenidos al entrenar sobre **MNIST** o **MNIST** transportado y testear sobre **USPS**.

MNIST, volver a entrenar un clasificador y testearlo en **USPS** da como resultado una mejora en la tasa de acierto, volvemos a utilizar la técnica $t-$SNE para ver como son las distribuciones luego del mapeo. La figura 5.14 muestra el resultado obtenido. Observamos que a diferencia de lo mostrado en la figura 5.10, ahora obtenemos que las observaciones de un mismo dígito en los dos conjuntos de datos se superponen, esto explica el por qué un clasificador entrenado en el conjunto **MNIST** transportado obtiene una mejor tasa de acierto.

Podemos observar de forma inmediata que ahora cada cluster asociado a cada dígito se solapa en mayor medida con su contraparte en el otro dataset, esto explica por qué vemos una mejora en el accuracy. Como contrapartida, también observamos que algunos puntos se mapean en una clase incorrecta. La figura 5.15 muestra algunas imágenes del dataset **MNIST** trasportado.

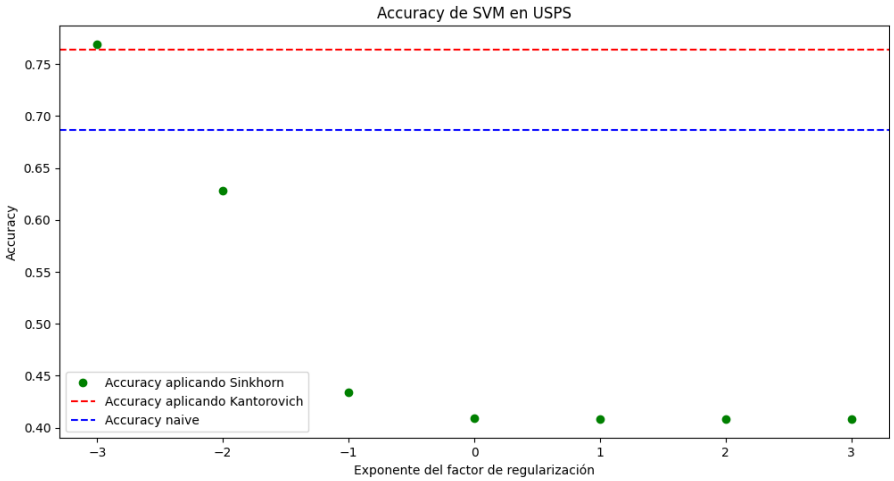


Figura 5.13: Tasa de acierto obtenido al variar el factor de regularización. La recta azul es el accuracy obtenido al no aplicar adaptación de dominio mientras que la recta roja es el accuracy que se obtiene aplicando el plan de transporte proveniente del problema de Kantorovich.

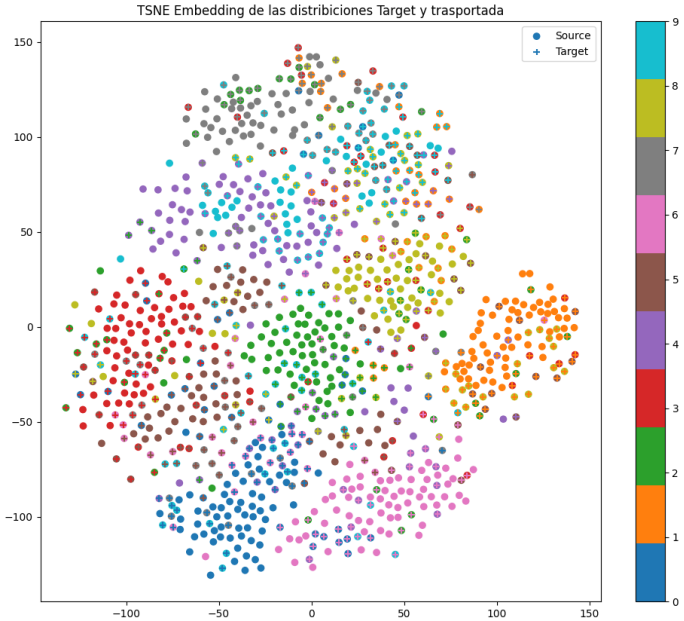


Figura 5.14: Proyección en el plano mediante la técnica $t-$SNE de los conjuntos **MNIST** transportado y **USPS**.

En la figura 5.15 podemos ver que algunas imágenes quedaron en una clase incorrecta, aunque en lineas genérales cada dígito quedo asociado a su clase. Una clara mejora a este procedimiento sería encontrar el transporte óptimo restringido a cada clase.

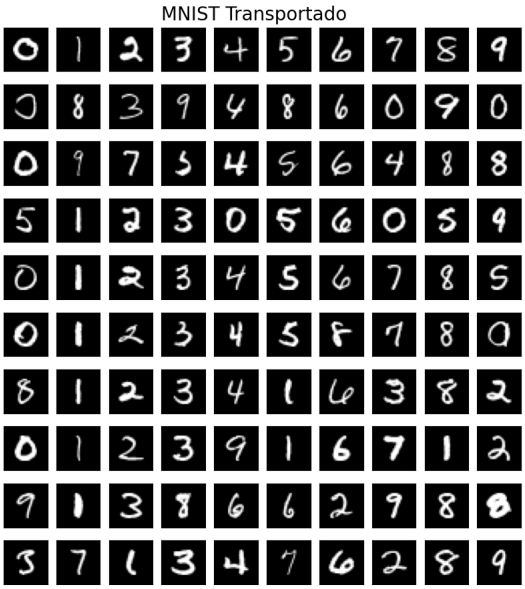


Figura 5.15: Algunos ejemplos del conjunto **MNIST** transportado. La columna indica en dígito. Vemos por ejemplo, que algunas observaciones de los dígitos 3, 5, 8 y 9 fueron transportadas a la clase del 0.

5.3. Transferencia de color

Para ejemplificar esta técnica transportaremos los colores de un atardecer a una foto de la Facultad de Ingeniería (FING en adelante). La figura 5.16 muestra ambas imágenes.

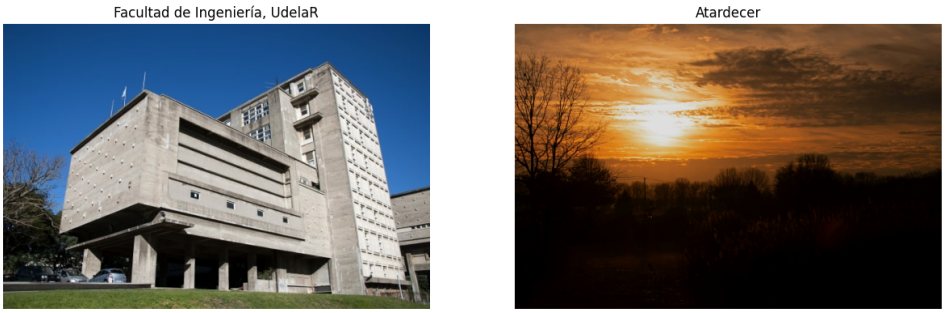


Figura 5.16: A la izquierda: Facultad de Ingeniería (FING). A la derecha: fotografía de un atardecer.

Convertimos cada imagen de dimensiones $(n_w, n_h, 3)$ en un matriz de dimensiones $(n_w \times n_h, 3)$, en donde cada columna representa un color primario $R - G - B$ (rojo, azul o verde) y cada fila representa un píxel de la imagen. Es decir, fijada una fila, las columnas nos dicen la intensidad de rojo, azul y verde del píxel asociado. En la sección 3.1 realizamos una explicación detalla

de este procedimiento.

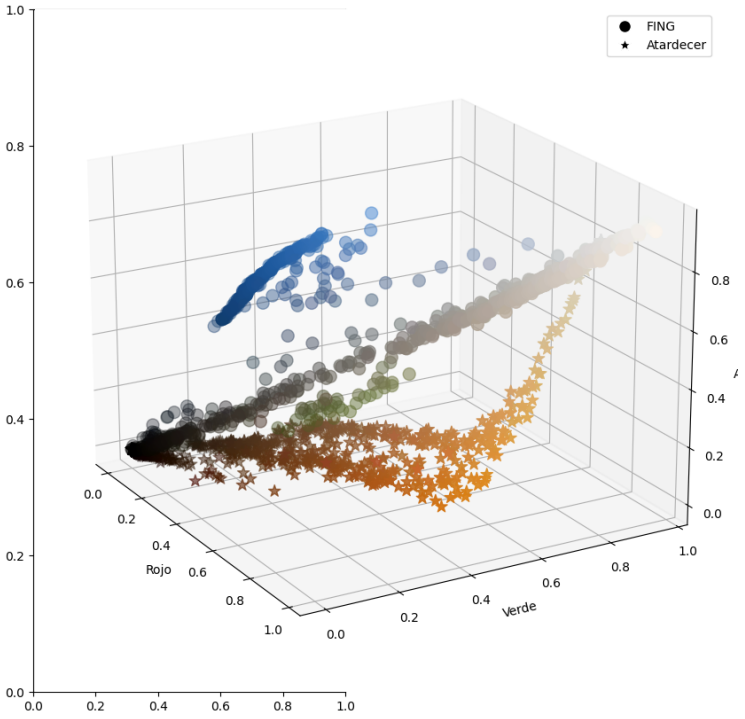


Figura 5.17: Cada dimensión representa un color. Los puntos circulares son los correspondientes a la fotografía de la FING mientras que las estrellas corresponden a la fotografía del atardecer. Es inmediato notar que la fotografía de la FING tiene una parte con cielo y por lo tanto en esta figura hay círculos con color celeste. Por otro lado, el atardecer tiene muchos píxeles con colores anaranjados.

La figura 5.17 es un scatterplot de esta matriz. De forma inmediata podemos comprobar lo que ya sabíamos: los píxeles del atardecer tienen mayoritariamente tonos negros y naranjas mientras que los píxeles de la fotografía de la FING tienen algunos píxeles azules (el cielo), otros verdes (el pasto) y otros grises (el edifico).

Utilizaremos la discretización Lagrangiana sobre los píxeles (ver sección 3.1) y como función de costo la distancia euclídea. La figura 5.18 muestra el resultado obtenido al aplicar el transporte óptimo obtenido del problema de Kantorovich y del problema regularizado con un factor de regularización $\epsilon = 10^{-1}$.

Figura 5.18: A la izquierda: Imágenes originales. A la derecha: arriba lo obtenido del problema de Kantorovich, abajo lo obtenido con el problema regularizado. El problema de transporte óptimo regularizado permite obtener una fotografía más suave, sin tanta saturación de color.

Es claro que la solución al problema regularizado da una mejor imagen (menos saturada) y esto no es de extrañar, dado que como ya mencionamos, la solución al problema regularizado es mas dispersa que la solución del problema de Kantorovich (ver sección 2.4).

Dado que al utilizar el transporte óptimo proveniente del problema regularizado con $\epsilon = 10^{-1}$ obtenemos una imagen menos saturada, es natural que nos preguntemos que efecto tiene en una imagen variar el factor de regularización.

Sabemos que cuando ϵ tiende a cero, el problema regularizado se convierte en el problema de Kantorovich, de donde obtendríamos una imagen similar a mostrada arriba a la derecha de 5.18. En contraste, a medida que aumentamos ϵ, la solución al problema regularizado tiende al producto de las medidas $\mu \otimes \nu$ y la solución es cada vez menos concentrada en una curva, lo cual permite ganar estabilidad. La imagen 5.19 muestra lo que sucede al aumentar el factor de regularización ϵ. Podemos observar que la imagen va 'desapareciendo' hasta un punto donde es indistinguible.

Figura 5.19: Efecto sobre la imagen al aumentar el factor de regularización ϵ.

En resumen, al aplicar el transporte óptimo al problema de transferencia de color es importante recordar que utilizar el problema regularizado puede dar mejores resultados siempre con un factor de regularización pequeño (entorno a 10^{-1}). En cambio, si el factor de regularización es demasiado grande, podemos llegar a perder toda la información de la imagen.

Conclusión

En este trabajo abordamos varias variantes del campo del transporte óptimo y la adaptación de dominio, estableciendo tanto una base teórica como algunas aplicaciones prácticas y ejemplos detallados.

El núcleo del trabajo se centra en el transporte óptimo, donde presentamos los fundamentos teóricos incluyendo las formulaciones de Monge y Kantorovich, y la distancia de Wasserstein. Estudiamos también la formulación dual y el problema de transporte óptimo regularizado, siendo este último usado frecuentemente en las aplicaciones. Además exploramos en detalle las relaciones entre las diferentes formulaciones.

Abordamos la implementación práctica del transporte óptimo con un enfoque en técnicas de discretización y algoritmos específicos, como el algoritmo Húngaro, que nos permite resolver el problema de asignación lineal y el algoritmo de Sinkhorn-Knopp, utilizado para resolver el problema de transporte óptimo regularizado. Además, discutimos aplicaciones prácticas del transporte óptimo y su integración con el aprendizaje automático, estudiando en detalle su utilización en el problema de la adaptación de dominio.

En cuanto a la adaptación de dominio, analizamos su motivación y proporcionamos definiciones básicas junto con una taxonomía de métodos. Estudiamos enfoques tanto supervisados como no supervisados, lo que permite la transferencia de conocimientos entre diferentes dominios con diferentes distribuciones y mejorar la eficacia de algunos modelos de aprendizaje automático en nuevos contextos. Estos métodos son particularmente relevantes en situaciones donde los datos de entrenamiento y los datos de prueba provienen de distribuciones diferentes.

Estudiamos también aplicaciones en conjuntos de datos reales para validar los conceptos teóricos presentados a lo largo del trabajo. Estas simulaciones incluyeron aplicaciones en clasificación de dígitos y la transferencia de color, demostrando la aplicabilidad y efectividad de los métodos de transporte óptimo

y adaptación de dominio en problemas prácticos.

Proponemos además un procedimiento para realizar adaptación de dominios cuando el modelo es una regresión lineal y los dominios source y target difieren en una rotación. Realizamos varias simulaciones para poner a prueba el procedimiento propuesto.

Como próximos pasos proponemos tres líneas de trabajo: profundizar en los problemas que involucran la distancia de Wasserstein, en particular en la geometría del espacio de Wasserstein y su vinculación con conceptos de la geometría de la información. Un ejemplo de esta linea de trabajo es Zemel y Panaretos (2016). Por otro lado, es de nuestro interés seguir explorando el problema de la adaptación de dominio con una regresión lineal, por ejemplo ampliando lo que realizamos componiendo la rotación con una traslación. Por último, una temática que surgió recientemente es el estudio de problemas de adaptación de dominio utilizando transporte óptimo en el contexto de series temporales (Painblanc et al. 2023).

Referencias bibliográficas

Arjovsky, M., Chintala, S., y Bottou, L. (2017). Wasserstein GAN.

Arun, K. S., Huang, T. S., y Blostein, S. D. (1987). Least-squares fitting of two 3-D point sets.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., y Peyré, G. (2014). Iterative Bregman Projections for Regularized Transportation Problems.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*.

Cao, H., Gu, H., y Guo, X. (2023). Feasibility of Transfer Learning: A Mathematical Framework.

Carlier, G. (2010). Optimal Transportation and Economic Applications.

Courty, N., Flamary, R., Tuia, D., y Rakotomamonjy, A. (2016). Optimal Transport for Domain Adaptation. *IEEE Transactions on pattern analysis and machine intelligence*.

Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transport distances.

Cuturi, M., y Doucet, A. (2014). Fast Computation of Wasserstein Barycenters.

Ferradans, S., Papadakis, N., Rabin, J., Peyré, G., y Aujol, J.-F. (2013). Regularized Discrete Optimal Transport. *International Conference on Scale Space and Variational Methods in Computer Vision*.

Galichon, A. (2016). *Optimal transport methods in economics*. Journal of Economics.

Genevay, A., Peyré, G., y Cuturi, M. (2017). Learning Generative Models with Sinkhorn Divergences.

Horn, R. A., y Johnson, C. R. (2013). *Matrix Analysis: 2nd edition*.

Hull, J. (1994). A database for handwritten text recognition research.

Kantorovich, L. (1942). On the transfer of masses (in russian). *Doklady Akademii Nauk*.

Kolkin, N., Salavon, J., y Shakhnarovich, G. (2019). Style Transfer by Relaxed Optimal Transport and Self-Similarity.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem.

Latorre, F., Liu, C., Sahoo, D., y Hoi, S. C. (2023). OTW: Optimal transport warping for time series.

Levy, B. (2014). A Numerical Algorithm for L2 Semi-Discrete Optimal Transport in 3D.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., y Zhang, G. (2020). Learning under Concept Drift: A Review.

Merigot, Q., y Thibert, B. (2020). *Optimal transport: discretization and algorithms*. hal-02494446f.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., y Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais.

Montesuma, E. F., Mboula, F. N., y Souloumiac, A. (2023). Recent Advances in Optimal Transport for Machine Learning.

Munkres, J. (1957). Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., y Lowe, R. (2022). Training language models to follow instructions with human feedback.

Painblanc, F., Chapel, L., Courty, N., Friguet, C., Pelletier, C., y Tavenard, R. (2023). Domain Adaptation of Time Series through Optimal Transport and Temporal Alignment.

Pan, S. J., y Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*.

Paty, F.-P., y Cuturi, M. (2020). Regularized Optimal Transport is Ground Cost Adversarial.

Peyre, G., y Cuturi, M. (2019). Computational Optimal Transport. *Foundations and Trends in Machine Learning*, *11*(5-6), 355-607.

Redko, I., Courty, N., y Tuia, D. (2019). Optimal Transport for Multi-source Domain Adaptation under Target Shift. *International Conference on Artificial Intelligence and Statistics*.

Rudin, W. (1991). *Functional Analysis (2nd ed.)* McGraw-Hill.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet†, D. J., y Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.

Sahiner, B., Chen, W., Samala, R. K., y Petrick, N. (2023). Data drift in medical machine learning: implications and potential remedies.

Santambrogio, F. (2017). Euclidean, metric, and Wasserstein gradient flows: an overview. *Bulletin of Mathematical Sciences*, *11*(7), 87-154.

Sinhorn, R., y Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices.

Staudt, T., Hundrieser, S., y Munk, A. (2022). On the Uniqueness of Kantorovich Potentials.

Taskesen, B., Shafieezadeh-Abadeh, S., y Kuhn, D. (2022). *Semi-discrete optimal transport: hardness, regularization and numerical solution*. Springer.

Thorpe, M. (2018). Introduction to Optimal Transport. University of Cambridge.

van der Maaten, L., y Hinton, G. (2008). Visualizing Data using t-SNE.

Villani, C. (2009). *Optimal Transport: Old and New*. Springer Verlag.

Wang, M., y Deng, W. (2018). Deep Visual Domain Adaptation: A Survey.

Zemel, Y., y Panaretos, V. M. (2016). Frechet means in Wasserstein space: gradient descent and procrustes analysis.

Zhang, J., Zhong, W., y Ma, P. (2021). *A Review on Modern Computational Optimal Transport Methods with Applications in Biomedical Research*. Springer International Publishing.